One Advisor for the Whole World? Cross-Country Evidence on Financial Advice from Large Language Models*

Claes Bäckman†

August 13, 2026

Abstract

Households around the world increasingly take portfolio advice from the same handful of large language models, replacing nationally segmented advice markets with model-based advice. Does that advice adjust to local circumstances? I pose the same portfolio problem to leading models across twenty-one countries, in different languages and with different local context. Advice is nearly uniform across countries, even though the explanations invoke local context, but advice varies considerably between models. Model identity accounts for 70 percent of the variation, compared with 6 percent for country. While model explanations tends to invoke local context if stated in the prompt, the advised risky share is barely affected by translating prompts or naming local financial instruments. In addition, country-level risky shares only respond weakly to differences in local financial fundamentals and to stated pension wealth. By contrast, the advice respond strongly to stated risk tolerance and investment horizon. The models appear to follow retail financial advice rather than what academic finance would prescribe, and only selectively incorporate the household's balance sheet when it is stated. The results suggest that large language models encode similar advice for investors around the world, and that the main determinant is which model the investor tends to ask for advice.

JEL Classifications: G11, G41, G51, D14, O33

Keywords: financial advice, household finance, portfolio choice, robo-advising, large language models

*I thank Joao Cocco, Lena Lieblich, Olga Balakina, Christine Laudenbach, Tobin Hanspal and Mikael Paaso, as well as seminar participants at Leibniz Institute for Financial Research SAFE for helpful comments. I gratefully acknowledge research support from the Leibniz Institute for Financial Research SAFE. The project received funding from Behavioral Finance e.V. The experimental design, regression specifications, primary hypotheses, and inference rules are pre-registered at the Open Science Framework (OSF) prior to the main data collection, accessible at https://osf.io/2ch5z.

†Leibniz Institute for Financial Research SAFE. Email: claes.backman@gmail.com

1 Introduction

Retail investors are increasingly turning to large language models (LLMs) for financial advice. Over half of US adults report they have used LLMs for personal financial guidance (Lloyds Banking Group, 2025; J.D. Power, 2025), and the share of U.S. adults that use artificial intelligence tools specifically for financial advice roughly doubled between 2024 and 2025 (Financial Health Network, 2025). Unlike nationally segmented advice markets where advisors tailor recommendations to local circumstances, financial advice from LLMs is concentrated among a handful of foundation models, trained overwhelmingly on U.S.-English text. This raises a question: What advice would LLMs give to investors in different countries, and would they tailor their recommendations to the features of the portfolio allocation problem that differ across countries?

I answer this question with a pre-registered experiment that poses the same portfolio problem to thirteen frontier LLMs across twenty-one countries: how should a 40-year old investor split the local-currency equivalent of $50,000 between a broad stock-market index fund and government bonds? The key finding is that while advice does vary across countries, the differences are modest and one-sided. Compared to the United States, every significant contrast is negative, ranging from 2.7 percentage points less equity in Mexico to 11 points less in Brazil. Seven of the twenty countries receive a significantly lower equity share than the United States, and averaged over all twenty countries the effect is -3.1 percentage points. I instead document a central role for model identity: the chosen model accounts for 70 percent of the variance in the primary cell against 6 percent for country. Mean advice spans 34 percentage points across the thirteen models against 12 points across the twenty-one countries. If there is only one optimal allocation, this disagreement between models implies that the average model must deviate from the optimum by at least the eleven-point standard deviation of the model means, a gap worth six basis points of the stake per year and roughly four times the median country gap.[1]

I vary the prompts to decompose the main results. Asking in the local language rather than in English lowers the recommended equity share by 1.3 percentage points within a country, and naming local instruments (e.g. a US 401(k) plan, a Swedish *tjänstepension*, local government bonds, and local index funds) moves it by less than one percentage point. But while these differences in prompts do not move the numbers, they do affect the *explanations* that the models provide: naming local instruments in the prompt raises local references in the prose by 23 percentage points. Local financial context appears in 18 percent of U.S. explanations but 30 to 78 percentage points more often in every other country. Overall, the models are aware of the local country context and reach for local context to explain the decisions, but this does not change the recommended portfolio allocation.

[1] Let w_m be model m's mean recommended share in the primary cell. For any optimum w^*, $M^{-1}\sum_m (w_m - w^*)^2 = \mathrm{Var}(w_m) + (\bar{w} - w^*)^2 \geq \mathrm{Var}(w_m)$, so the certainty-equivalent loss (13) averaged across the thirteen models is at least $\frac{1}{2}\gamma\sigma^2 \mathrm{Var}(w_m)$. The standard deviation of the model means is 0.111, so at $\bar{\gamma} = 3$ and $\sigma = 0.18$ the floor is 6.0 basis points of the liquid stake per year against a median priced country gap of 1.6.

Why is the advice almost invariant to the language of the question? One explanation is that the personal financial advice is similar in different languages, so that the models are reading materials that just do not differ much across countries. Essentially, under this explanation financial advice in the German source materials is similar to financial advice in the Chinese source materials, and models drawing on German and Chinese material would answer similarly. The invariance then reflects an invariance in the advice itself in different countries. The alternative explanation is that the models solve the problem in a shared, English-anchored representation regardless of the language of the prompt, and render only the explanation in the local language. Essentially, under this explanation the models draws on a shared knowledge base to come up with the number but renders the explanation in German and Chinese. Translating the prompt changes the words of the question and the answer without reaching the computation that sets the number. Then the invariance reflects how the models work rather than any invariance in the advice itself. This is what the interpretability literature finds for multilingual models generally, which solve tasks in a shared, English-leaning concept space rather than retrieving language-keyed knowledge (Wendler, Veselovsky, Monea and West, 2024; Tang, Luo, Huang, Zhang, Wang, Zhao, Wei and Wen, 2024; Wu, Yu, Yogatama, Lu and Kim, 2025; Lindsey, Gurnee, Ameisen, Chen, Pearce, Turner, Citro et al., 2025). My evidence is consistent with this also happening for financial advice, as asking in the local language with local contect changes the recommended share by 0.7 percentage points but increases local references in the explanation by 23 percentage points.[2]

I ask two additional questions that further illustrates how models use local context and the scenario stated in the prompts. One question asks whether an individual should pay back on their 4% mortgage or invest in the stock market. The average fraction to invest is close to 70 percent for the mortgage question, compared to 54 percent for the benchmark question described in the last paragraph. The cross-country differences are also large, where the fraction to invest falls an average of 14 percentage points below the U.S. baseline across the twenty countries compared to -3.1 percentage points for the benchmark question. The additional second question asks how an investor should protect the real value of their savings against inflation. This latter question asks both what fraction of the local-currency equivalent of $50,000 to invest, and the primary instrument to use to protect savings from inflation. The recommended amount to invest moves varies little across countries (-2.6 percentage points). However, the recommended instrument changes depending on local context. Equities are recommended in over 93 percent of responses in fourteen countries, but in Brazil and Turkey, the models recommend inflation-linked bonds instead, from 87 percent of Brazilian responses to 66 percent of Turkish ones, with gold taking a further 16 percent in Turkey. The limited variation in the equity share in the baseline question

[2] An illustrative case from the explanation text: I validated the Swedish terms by myself, and noticed that one Swedish response recommends a "så kallad glidväg" ('so-called glide path'), a direct translation of the English term that I have never heard in Swedish. A Google search for "glidväg finans" confirms that the term is not used for financial terms in Swedish, although ironically, the Google AI overview provides an explanation for the term as if it existed in Swedish. I did not see a reference to *generationsfond*, an established Swedish term for the product category, in the explanation text.

therefore does not reflect that the models are insensitive to the country. If the question supplies a local quantity that is important for advice, like the mortgage rate, then the advice moves. However, the models does not supply the local fundamentals themselves without prompting, which illustrates the importance of good prompting (Choukhmane, de Silva, Lin and Akuzawa, 2026). On the risky share itself, the models behave less like twenty-one national advisors and more like one advisor answering in twenty-one languages.

One can fairly ask at this point what the cross-country level and variation in finanical advice should actually be. We can make a good case for uniform advice across countries by noting that basic finance theory, like Merton (1969), has the optimal risky as a function of the return premium, volatility and risk aversion. Under a global portfolio, the market return and the volatility are the same across all countries, and since the baseline question does not supply a risk aversion level, a model that assumes that this is constant is not making a mistake, per se. What would motivate an advisor to offer different advice in different countries for otherwise identical investors, however, are the local safe rates, currency risk, and above all the pension systems, which shift the background wealth against which the liquid portfolio is chosen. There are differences across countries in many factors that may shift the optimal risky share, and since the baseline results suggest that there is some variation across countries, I now turn to the question of what the models are responding to. Are the models responding to quantities and variables present in standard portfolio choice models?

To do so, I develop a country-specific normative benchmark in two forms to understand much variation country fundamentals should generate under standard portfolio benchmarks. I begin with a classic standalone Merton portfolio rule and an continue with an extension with background wealth in the form of pension wealth and human capital. The benchmarks allow me to calculate an equity share based on local fundamentals in each country. I find that the advice does not track what local fundamentals prescribe. Under identical risk aversion, optimal equity shares under the standalone benchmark, evaluated in local currencies, range from 0.21 in Turkey to 0.53 in South Korea. Yet mean advice spans only 0.46 to 0.58, a much smaller range. Allowing risk aversion to vary with country-level preference measures from Falk, Becker, Dohmen, Enke, Huffman and Sunde (2018) does not change this result. However, since this is a normative benchmark that is difficult to defend as representing the optimal portfolio share in each country, I use two further exercises to examine what correlates with the country-level variation in the advice. I relate the estimated country fixed effects to country-level risk-taking from the Global Preferences Survey (Falk et al., 2018) and to financial literacy from the S&P Global FinLit Survey (Klapper, Lusardi and van Oudheusden, 2015). However, I find only a weak relationship between either measure and the average equity share a country receives. Overall, the advice is more uniform than standard portfolio benchmarks would predict, and the models do not seem to read local fundamentals

unprompted.

In the final part of the paper, I vary investor characteristics by inserting statements of risk tolerance, pension generosity, or investment horizon into otherwise identical English prompts for each country. The goal is to better understand what the models do respond to on the investor level. Each treatment targets a theoretically motivated part of the portfolio problem. Stated risk tolerance should raise the recommended equity share, and I find that it does: stated risk tolerance raises the recommended equity share by 46 percentage points. This result is similar in Ross and Lo (2026), Li, Xue, Zeng and Zhao (2026) and Foltyn and Olsson (2026). Next, I include a statement about pension coverage in retirement, which should raise the equity share if the model integrates the investor's balance sheet. However, the stated pension only lowers the recommended share by about a point for the risk-averse investor and raises it by under three points for the risk-tolerant one. This is not because models ignore pensions, as nearly every explanation text from the models mentions existing pensions. Finally, the stated horizon should not move the equity share at all under CRRA preferences with i.i.d. returns, the classic Merton–Samuelson irrelevance result. However, I find that it does: a twenty-year rather than one-year horizon raises it by 40 points, again similar to the results in Ross and Lo (2026) and Li et al. (2026). This a response that time-varying expected returns could rationalize even though the static benchmark cannot.

I use these results to infer the candidate model of financial advice that the LLMs are using. Under a standard frictionless Merton planner, the optimal share is horizon-invariant. This is clearly rejected in all countries. The pension null excludes a lifecycle planner with background wealth, for whom a bond-like pension claim must raise the equity share of the liquid portfolio. Instead, the advice patterns fits well with "popular financial advice", as documented in Canner, Mankiw and Weil (1997) and Choi (2022): a risk questionnaire keyed to stated tolerance and horizon, rather than a portfolio rule over the household's balance sheet. This is also consistent with the pension null-result in the investor-arm of the study.

In conclusion, this paper characterizes the supply side of a newly integrated market for financial advice from LLMs. Financial advice was a nationally segmented service industry, walled in by national regulation, by the trust relationships that a credence good requires, and by language. While human advisors already gave one-size-fits-all advice, with advisor fixed effects dominating client characteristics (Foerster, Linnainmaa, Melzer and Previtero, 2017), LLMs are further pushing financial advice toward a one-size-fits-all structure by globalizing the distribution of advice. Empirically, I show that advice compressed toward a common level at or a few points below the U.S. baseline for most countries, localization responds to the country name but barely to the question's language or to explicitly named local instruments, and the advice seems to follow a retail risk questionnaire rather than a portfolio rule over the household's balance sheet. This paper shows that large language models can answer questions about financial advice in every language at negligible marginal cost, outside the regulatory perimeter

that bound financial advice from humans.

2 Literature

This paper contributes to three literatures. Most closely related to this paper is Choukhmane et al. (2026), who provide the first systematic quantitative account of LLM financial advice. They elicit prompts from a representative U.S. sample, simulate the life-cycle paths that following the advice would generate, and confront those paths with the prescriptions of a calibrated life-cycle model. They find that the advice of current models is moderately risk averse, broadly consistent with the broad principles of life-cycle theory yet reliant on round-number heuristics and varying with gender and financial literacy. Compared to this paper, both its normative benchmark and its experimental variation are within a single country, so it cannot ask whether the advice is calibrated to the institutions, asset markets, and preferences of the household being advised. Moreover, it reduces each response to a numeric allocation and discards the model's stated reasoning by construction, instructing the model to return allocations without explanation, whereas I retain and code the explanation text to ask whether the reasons the model gives track local fundamental.

A closely related and contemporaneous paper, Ross and Lo (2026), reaches the same conclusion about the advice rule as this paper from a different methodological direction. They draw a thousand synthetic client profiles with near-orthogonal characteristics, collect allocations from a single vendor's model family, and fit interpretable surrogate models to the input-output mapping. They find evidence for what they call heuristic collapse: self-reported risk tolerance accounts for most of the explained variation in the allocations while age, income, horizon, and liquidity contribute marginally. That their population-level surrogate approach and my treatment-based experiment recover the same advice rule strengthens both readings. My contribution differs on two margins. First, I take the analysis to the country level. Their synthetic profiles are drawn from a single vendor's model family and hold the investor's country fixed, so their design cannot ask whether the advice adjusts to the institutions, asset markets, and preferences of the country the investor lives in, whereas I vary the country and the language of the prompt across twenty-one countries and thirteen models from six vendors. Second, I build a country-calibrated normative benchmark that prices the failure and locates its incidence across countries, whereas their design declines a normative benchmark by construction and shifts the evaluative criterion from output accuracy to input sensitivity. Third, I isolate one capacity channel by stating pension generosity in the prompt itself, rather than reading it off a jointly varying profile.

To the economics of robo-advising and algorithmic financial advice, it adds evidence on whether the newest generation of algorithmic advisors localizes its recommendations. D'Acunto, Prabhala and Rossi (2019) show that algorithmic advice improves diversification and reduces behavioral biases, Rossi and

Utkus (2024) document substantial welfare gains from robo-advising, and D'Acunto and Rossi (2023) survey the broader evidence that robo-advisors move household behavior closer to normative prescriptions, while Reuter and Schoar (2024) emphasize that demand- and supply-side constraints jointly determine who accesses and benefits from advice. LLMs differ from earlier robo-advisors in two ways that motivate this paper. They are general-purpose rather than domain-restricted, so their advice is not constrained to a particular asset universe, and they are deployed in essentially every language and country simultaneously, so any country-specific bias in their training data propagates globally. Whether LLMs deliver the gains documented for first-generation robo-advisors depends in part on whether their advice is appropriately localized.

To the literature on cross-country differences in household financial behavior, this paper adds a new margin: the advice households receive, as distinct from the choices they make. Guiso and Sodini (2013) and Christelis, Georgarakos, Jappelli and van Rooij (2015) document substantial cross-country variation in equity-market participation that is not fully explained by income, wealth, or institutional access, the Global Preferences Survey (Falk et al., 2018) provides validated country-level preference measures that have been linked to savings and investment behavior, and van Rooij, Lusardi and Alessie (2011) show that financial literacy is a strong predictor of equity-market participation.

To the literature on how LLMs encode demographic and social stereotypes (Bolukbasi, Chang, Zou, Saligrama and Kalai, 2016; Navigli, Conia and Ross, 2023), this paper adds both a new treatment and a decomposition. Fuster, Goldsmith-Pinkham, Ramadorai and Walther (2022) show that the introduction of machine-learning credit-scoring tools changes the distribution of mortgage approvals across racial groups, and Foltyn and Olsson (2026) document gender-based gaps in LLM portfolio advice. I shift the treatment from demographic identity to country and language, a dimension with first-order policy relevance given the borderless deployment of LLM-based advisory tools and the country-specific structure of pension and tax systems. Methodologically, I decompose the country-and-language effect into language, country-naming, and local-framing components, separating mechanisms that are conflated in any single-prompt comparison.

3 Experimental Design

The main experiment poses three financial questions to thirteen frontier LLMs in twenty-one countries, varying the language of the question, the country reference in the question, and the local financial context provided in the question. This section describes the treatment dimensions, the three outcome questions, the identification, the stated-investor arm that varies the investor rather than the country, the country selection, and the translation workflow. Appendix Table A1 lists the prompt language, the local-currency investment amount, and the full Version C vocabulary for all twenty-one countries that I describe in more

detail in this section. Appendix B documents the technical implementation of the data collection in more detail.

3.1 The Three Questions

Each LLM is asked three financial questions. The primary outcome is Question 1, which asks the investor to allocate the local-currency equivalent of$50,000 between a broad stock-market index fund and government bonds over a five-year horizon. The outcome is the fraction allocated to equities. Question 1 is the outcome on which all primary hypotheses are tested.

Question 2 asks whether to invest savings in a stock-market index fund or to use them to pay down an existing 4% mortgage with twenty years remaining. The outcome is the fraction allocated to the index fund. Q2 has a sharper normative benchmark than Q1, as mortgage prepayment is a risk-free investment whose return is pinned down by the quoted rate. A pre-specified rule assigns any country whose prevailing mortgage rate exceeds twice the prompted 4% to a premise-acceptance group—expected to include Brazil, Turkey, Mexico, India, and South Africa, and possibly Poland—whose cells test whether models flag the 4% rate as unusual for the local environment or simply condition on it (in the end, all conditioned on it). A fixed 4% mortgage is unusual, not impossible, where prevailing rates are far higher, and its prepayment return is pinned by the contract rate regardless.

Question 3 asks how to protect the real value of $50,000 sitting in a savings account that earns 1% nominal while annual inflation is 3%. The outcome has two parts: a numeric fraction-to-invest and a categorical `primary instrument` (equities, inflation-linked bonds, real estate, gold, other).[3]

3.2 Country Selection

The country set is chosen to span meaningful cross-country variation in pension generosity, equity market development, inflation history, and measured preferences, while keeping translation risk low. To begin, I applied one screening constraint to the country list: the preference-calibrated benchmark in Section 5 requires each country's measured risk preferences, so every country must be covered by the Global Preferences Survey (Falk et al., 2018). This requirement excludes otherwise natural candidates such as Denmark. The United States serves as the baseline because U.S.-English text is the dominant component of LLM training corpora for most models, making it the natural reference category for the country contrasts. Five further English-prompting countries (Australia, Canada, India, South Africa, and the United Kingdom) add observations at essentially no translation risk while spanning very dif-

[3]Question 3 is the question where, *ex-ante*, it was more likely that the models would refuse the premise and not give a numerical answer. To guard against this, a pre-specified contingency promotes the refusal indicator to the primary Q3 outcome if refusal rates exceed 10% in any country-model. However, the refusal rate was 0% across 37,021 model calls, or 273 country x model cells, and this did not turn into a practical concern. Several models noted that the rate was unrealistically low for that country (most often for Turkey), but none refused to answer.

ferent pension institutions, from compulsory superannuation to a replacement rate among the lowest in the panel. Because these countries share the baseline's language, they isolate country context from language: any U.S.–UK difference, for example, reflects the country reference rather than the language of the question.

The fifteen non-English countries are chosen to spread the fundamentals the benchmark of Section 5 prices: pension generosity, equity-market development, inflation history, and measured preferences. They run from high-literacy European systems with structurally different occupational pensions (Sweden, Germany, the Netherlands, Switzerland) through further European variation (France, Spain, Italy, Poland, Finland) to the Asian markets (China, Japan, South Korea), the high-inflation economies where the inflation question is most informative (Brazil, Turkey), and Mexico. Appendix Table A1 gives each country's language, amount, and local instruments.

3.3 Prompt Structure and Version Design

Each question prompt varies in three dimensions. The first is the country–language pair: I prompt six countries in English (the United States, the United Kingdom, Australia, Canada, India, and South Africa) and fifteen in the local language (Sweden in Swedish, Finland in Finnish, Germany in German, France in French, Spain in Spanish, Japan in Japanese, Brazil in Portuguese, the Netherlands in Dutch, Italy in Italian, Switzerland in German, Poland in Polish, South Korea in Korean, China in Chinese, Mexico in Spanish, and Turkey in Turkish). The second is the investor's stated age (30, 40, or 55), varied for the primary equity-share question but held fixed at 40 for the two secondary questions.

The third is the prompt version, which determines how the country and the local financial context are conveyed and which provides the within-country variation used to decompose the country effect into language, country-naming, and local-framing components. Every country is observed in Versions B and C, and each of the fifteen non-English countries is additionally observed in Version B-cross, so the six English-prompting countries have two versions and the fifteen non-English countries three.[4] For the six English-prompting countries Versions B and C are already in English and differ only in whether the named instruments are generic or local, so the language step isolated by Version B-cross does not arise for them.

Version B is translated into the local language and states "I live in [country]" but otherwise uses the generic instrument vocabulary of the U.S. baseline ("a broad stock market index fund," "government bonds," "a standard pension plan"). The investment amount in Version B is the round local-currency

[4]The version lettering follows the pre-registration. Version A denoted a translated prompt that named no country. It was dropped from the design before collection because it could not separate the effect of the question's language from the style of any particular translation, and the comparison of Versions B and B-cross identifies the language effect more cleanly, holding the country reference fixed. The labels of the surviving versions are kept unchanged so that the paper matches the registered design.

equivalent of $50,000 (e.g. 500,000 kronor in Sweden, 45,000 euros in Germany, 40,000 pounds in the United Kingdom), written in the native local number format ("500 000 kr," "45.000 Euro").

Version B-cross retains the country reference ("I live in [country]") and the local-currency investment amount of Version B but reverts the question to English. It is collected only for the fifteen non-English countries.

Version C is also translated into the local language but additionally replaces the generic financial vocabulary with country-specific equivalents—local index funds (e.g. an index fund via Avanza or Nordnet in Sweden, eMAXIS Slim All-Country Equity in Japan), local government bonds (*Svenska statsobligationer*, JGBs, Tesouro Direto), and local pension references (*tjänstepension*, bAV, PER, iDeCo).

I use these three contrasts to identify the components of the country effect. The cross-country comparison of Version B fixes language to the local language and identifies the joint country-and-language effect relative to the U.S. baseline. The within-country comparison of Version B-cross and Version B holds the country reference fixed and changes only the question's language, identifying the language effect. The within-country comparison of Versions C and B holds both country and language fixed and changes only whether local instruments and pension names are named explicitly, identifying the local-financial-context effect.

3.4 The Stated-Investor Arm

The versions above vary where the investor lives and in what language she asks. A final arm varies the investor herself. Version B$\gamma\rho\tau$ takes the English Question 1 prompt at age 40 and inserts statements of the investor's risk tolerance, pension generosity, or investment horizon. The versions are otherwise identical to Version B in the six English-prompting countries and to Version B-cross in the fifteen others, with the local-currency amounts unchanged.

The two risk-tolerance statements are length-matched qualitative sentences: "I am comfortable with substantial fluctuations in my portfolio value, including potential short-term losses, if it means higher expected returns over time" versus "I am uncomfortable with large fluctuations in my portfolio value and would prefer to limit potential losses, even if it means lower expected returns." The two pension statements are: "I will receive a substantial pension when I retire, projected to replace most of my pre-retirement income" versus "I will receive only a small pension when I retire, covering a modest fraction of my pre-retirement income". The statements are intentionally institution free. The horizon treatment replaces the phrase "over a five-year horizon" with a one-year or a twenty-year horizon and inserts no other statement. The risk and pension statements are fully crossed at the default five-year horizon (four cells), while the two horizon cells stand alone. The standard Version B and B-cross rows, whose prompts carry no statement and the default horizon, serve as the no-treatment reference.

I pre-registered three hypotheses over this arm as secondary confirmatory tests, each a single pre-specified coefficient reported without multiplicity correction and with the same wild-cluster-bootstrap inference as the primary regressions: stated risk tolerance raises the recommended equity share (with the directional prediction of the Merton rule), stated pension generosity moves it (two-sided, since the standalone reading of the question predicts no effect and the balance-sheet reading predicts a positive one), and the stated horizon moves it (tested against the CRRA benchmark of exact irrelevance).

3.5 Translation Workflow

I produced the Version B and Version C question blocks for the fifteen non-English countries with the DeepL API through an automated pipeline that translates only the Question block and protects the pinned financial terms and locked currency amounts. Machine output passed an automated term-protection check and a full back-translation, but the operative quality gate was native-speaker review of the B and C blocks in all fifteen translated countries, with a documented two-vendor LLM cross-check standing in for the English-prompting countries. Before the pre-registration was filed the full grid passed a consistency audit with zero failures and the translations were frozen for collection. Appendix B.2 documents the pipeline mechanics, the two automated checks and their failure modes, and the review-and-correction protocol.

Before the pre-registration was filed the full grid passed a consistency audit of every assembled prompt with zero failures, based on a LLM model-based adversarial review of the translated blocks. I also did a complete read of all ninety unique question blocks to ensure consistency across languages. The translations were then frozen for collection. Appendix B.2 documents the pipeline mechanics, the two automated checks and their failure modes, and the review-and-correction protocol.

3.6 Models, Repetitions, and API Parameters

The experiment queries thirteen frontier LLMs from six vendors via the OpenRouter unified API: four Anthropic models (Claude 4.5 Haiku, Claude 4.6 Sonnet, Claude 4.6 Opus, Claude 4.8 Opus), two OpenAI models (GPT-5-mini, GPT-5.5), three Google models (Gemini 3 Flash Preview, Gemini 3.5 Flash, Gemini 3.1 Pro Preview), two DeepSeek models (DeepSeek V4 Flash, DeepSeek V4 Pro), xAI's Grok 4.3, and Mistral's Mistral Medium 3.5. DeepSeek is included as a non-U.S. vendor whose models are primarily trained on Chinese-language data, and Mistral as a European (French) vendor, so that the panel is not composed solely of U.S.-headquartered firms.

Each cell consists of one country, one version, one question, one age, one model. Each cell is queried 50 times at temperature 0.7 with no system prompt and a maximum response length of 4,000 tokens. The unit of inference is the model rather than the repetitions, so fifty repetitions suffice. Additional

repetitions reduce only the within-cell sampling term, which is already small relative to between-model heterogeneity. The exact model slug returned by the API is logged with every raw response so that within-slug version drift can be monitored.

In total the design comprises 267,150 queries, of which 185,250 form the country/language experiment and 81,900 the stated-investor arm. Appendix Table A2 gives the accounting by arm, version, and question. The experiment is pre-registered at the Open Science Framework (OSF) after a staged pilot of roughly 6,800 queries passes three pre-specified success criteria: a JSON parse rate of at least 95 percent per model, refusal rates below pre-specified ceilings in every country-model cell, and a between-model dispersion of the country contrast no larger than 0.06.

3.7 Data Collection and Parsing

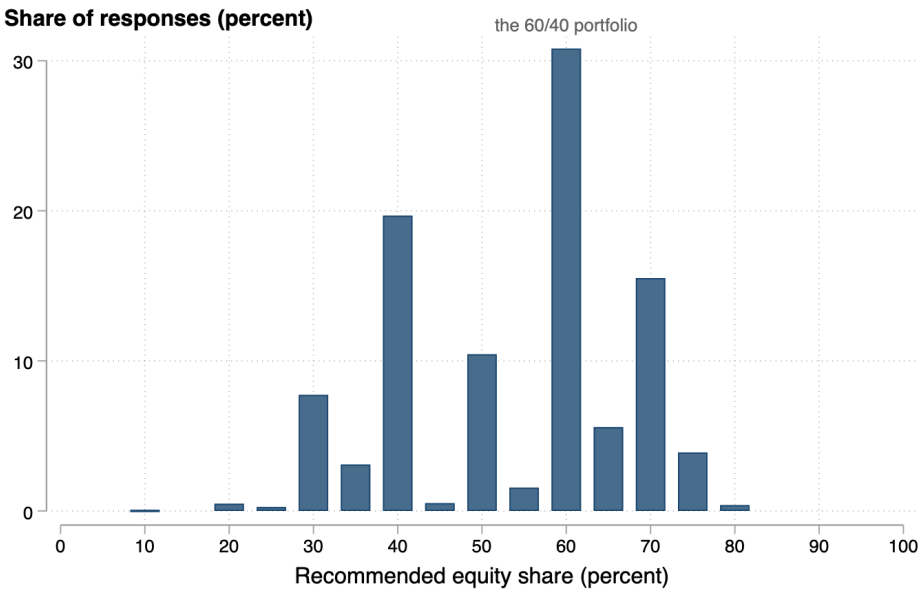
Collection proceeded in two stages, an exploratory pilot followed by the pre-registered main collection. A staged pilot of roughly 6,800 queries completed in late June 2026 exercised the full pipeline end to end and passed the three pre-specified success criteria, with a JSON parse rate near 99.6 percent, no refusals, and no truncated responses. The pilot preceded and informed the pre-registration, which was filed on July 2, 2026 with the prompts frozen. The main collection ran on July 3–4, 2026, submitting all 267,150 queries through the OpenRouter API within a two-day window, well inside the four-week bound the design places on a collection phase to limit drift in the underlying model weights. The two pre-registered collection-robustness cells—temperature bookends at 0 and 1.0 and a Version B variant that states the investment amount in U.S. dollars—added 8,450 queries on July 6, 2026.

The main run finished with exact coverage of the design grid, every country–version–question–age–model cell reaching its fifty repetitions with no duplicates and no unresolved API errors. A JSON-first parser with conservative fallbacks recovered a valid numeric recommendation from 99.9 percent of responses, the parsing strategy behind each observation is recorded in the data, and the remainder are excluded, together with responses that exhausted the 4,000-token budget, under the pre-registered truncation rule. The collection produced no refusals in 267,150 responses, so the pre-specified Q3 contingency, which would have promoted the refusal indicator to the primary Q3 outcome had any country–model cell refused more than 10 percent of the time, was not triggered. Model identity was stable throughout: every mismatch between the requested and the returned model identifier was the provider's resolution of a version alias to its dated snapshot, and no model rotated versions mid-run.

4 Main Results

This section describes the main results of the paper. Except for the two secondary questions taken up in Section 4.4, the regressions in this section take the recommended equity fraction on Question 1 as the

Figure 1: Distribution of recommended equity shares in the primary cell



Notes: Share of responses at each recommended equity share on the primary cell (Question 1, Version B, age 40, $N = 13{,}639$), pooled across the twenty-one countries and thirteen models. Every response is a multiple of five percentage points, so the display is a discrete bar chart rather than a histogram. No response falls below 10 or above 80 percent.

outcome. Every regression includes model fixed effects so that each contrast is identified within model, and clusters standard errors at the model level. Because thirteen clusters are too few for reliable asymptotic inference, every reported p-value comes from a wild cluster bootstrap with Webb six-point weights and 9,999 replications, and the twenty primary country contrasts carry a Holm step-down correction as a single family to account for multiple hypothesis testing. Within each comparison the queries pose the same investment problem, so the coefficients measure the causal effect of the prompt's country, language, and framing content on the advice.

4.1 Descriptive Patterns

I begin by documenting several features of advice from LLMs that organize the results that follows: financial advice from LLMs is bunched at round numbers, the advice occupies a narrow band in every country, and the residual variation across repetitions is small. I focus on the answers to the baseline cell, which asks about the share allocated to equity versus government bonds in the local language in the local currency, for an investor of age 40.

Figure 1 plots the distribution of the recommended equity share, pooled across countries and models. Every single response is a multiple of five percentage points, and 85 percent are multiples of ten. The modal recommendation is 60 percent equity, given in 31 percent of responses followed by 40 percent (20 percent of responses), 70 percent (16 percent), and 50 percent (10 percent). The support is strikingly narrow: no response in 13,639 allocates less than 10 or more than 80 percent to equities, so the corner

Table 1: Summary statistics by country on the primary cell

Country	Mean	SD	p10	Median	p90	*N*
Brazil	0.46	0.14	0.30	0.40	0.70	650
Italy	0.48	0.14	0.30	0.50	0.70	650
Poland	0.50	0.14	0.30	0.50	0.70	650
France	0.51	0.14	0.30	0.60	0.70	649
China	0.52	0.10	0.40	0.50	0.65	649
Switzerland	0.53	0.14	0.30	0.60	0.70	649
Spain	0.53	0.12	0.40	0.60	0.70	648
Netherlands	0.53	0.14	0.30	0.60	0.70	649
Germany	0.53	0.14	0.30	0.60	0.70	650
Turkey	0.54	0.11	0.40	0.60	0.65	650
Mexico	0.54	0.14	0.30	0.60	0.70	650
Finland	0.55	0.15	0.35	0.60	0.75	648
India	0.55	0.12	0.40	0.60	0.70	650
Canada	0.56	0.13	0.40	0.60	0.70	650
South Africa	0.56	0.12	0.40	0.60	0.70	648
Sweden	0.57	0.13	0.40	0.50	0.75	649
United States	0.57	0.13	0.40	0.60	0.70	650
Australia	0.58	0.12	0.40	0.60	0.70	650
United Kingdom	0.58	0.13	0.40	0.60	0.75	650
Japan	0.58	0.14	0.35	0.60	0.70	650
South Korea	0.58	0.11	0.40	0.60	0.70	650

Notes: Distribution of the recommended equity share by country on the primary cell (Question 1, Version B, age 40), pooled across the thirteen models with fifty repetitions per country–model cell. Countries are ordered by mean. *N* falls slightly below 650 where responses without a parsed numeric recommendation, or truncated at the token limit, are excluded under the pre-registered rules (Section 3.7).

allocations that the benchmark of Section 5 sometimes prescribes are never given.

The models thus answer a continuous allocation problem with a round number — 60/40 and its neighbors. This also happens to be the most common target share from retail portfolio advice — a first sign of the risk-questionnaire interpretation that Section 6 later develops. An alternative explanation is that the models are following a 100-minus-age rule, which would also predict a 60/40 allocation. The modal response for age 30 is also 60, although the age 55 modal response is 40 instead of 45 (see Figure A1 in Section 4.5). Regardless, it appears from the raw model output that the models are following either the popular financial advice or a rule of thumb instead of a measure that would provide a continuous answer.

Turning to summary statistics, Table 1 shows that mean advice spans a band of twelve percentage points, from 0.46 in Brazil to 0.58 in South Korea, with the United States at 0.57 near the top of the range. The band is narrow relative to the dispersion within any single country: the typical within-country standard deviation is 0.13, so the twenty-one distributions overlap almost completely and the full range of country means is no wider than one within-country standard deviation.

Appendix Table A3 shows that which model a household consults matters an order of magnitude more than which country it consults from. The table reports mean advice by model: the panel spans 0.37

(Gemini 3.1 Pro) to 0.71 (Claude Haiku 4.5), a range of 34 percentage points, nearly three times the country range. A variance decomposition shows that model identity accounts for 69.5 percent of the total variation in the primary cell, the country for 5.7 percent, their interaction for 9.5 percent, and sampling variation within country–model cells for the remaining 15.3 percent. These shares are computed over the equal-weighted panel of thirteen models and describe the dispersion across this non-random set, not the dispersion a representative household faces, which would depend on the models' market shares.

Interestingly, the model ordering follows neither vendor nor scale: the four Anthropic models sit between 0.58 and 0.71, all above the pooled mean. Google's models span nearly the full panel range, with its flagship reasoning model the most conservative advisor in the panel and its two Flash-tier models twenty points higher. These facts motivate two choices maintained throughout the paper: every regression identifies country contrasts within model, and inference clusters at the model level, treating the thirteen models rather than the quarter-million responses as the effective sample.

4.2 Country effects on the primary question

I now turn to the main country estimation, testing whether financial advice from LLMs is the same across countries. The primary specification estimates country fixed effects on the pre-registered identification cell, Version B of Question 1 at age 40,

$$y_{imc} \;=\; \alpha_m + \gamma_c + \varepsilon_{imc}, \tag{1}$$

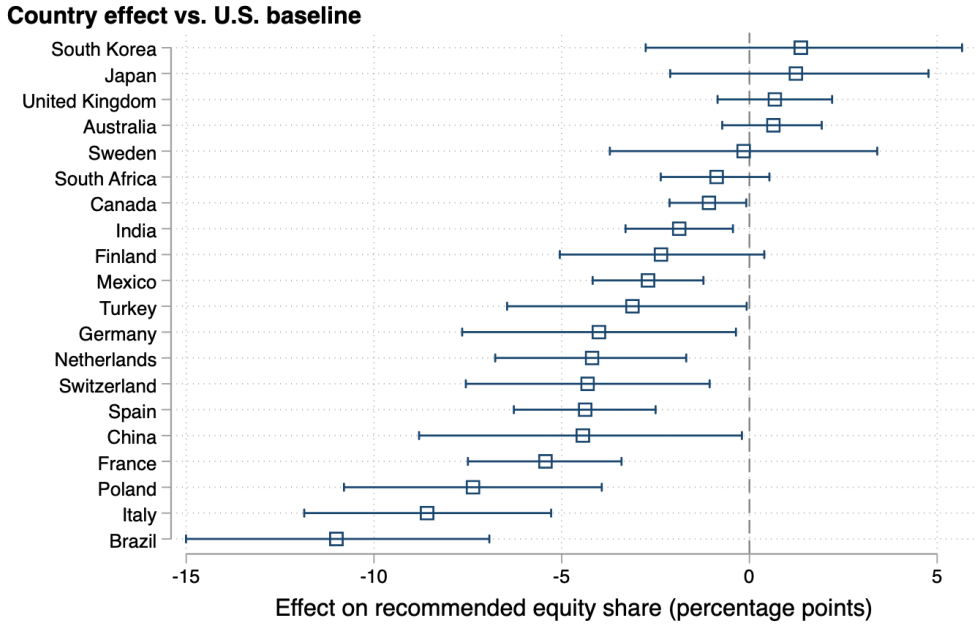
where y_{imc} is the equity fraction recommended in repetition i by model m for country c, the α_m are model fixed effects, and the γ_c are deviations from the omitted U.S. baseline.

The main finding is that advice is not the same everywhere. The joint wild-bootstrap test rejects equality of advice across the twenty-one countries with $p = 0.006$.[5] Figure 2 plots all twenty contrasts, and Appendix Table A7 reports each with its wild-bootstrap and Holm p-values. Thirteen of the twenty contrasts clear the wild-bootstrap threshold before correction, and seven survive the Holm step-down as a single family: Brazil at −11.0 percentage points, Italy at −8.6, Poland at −7.4, France at −5.4, Spain at −4.4, the Netherlands at −4.2, and Mexico at −2.7. Mean advice in the U.S. baseline cell is an equity share of 0.57, so the estimated effects place every country mean between 0.46 and 0.58. The design's minimum detectable effect for a single country contrast is roughly four percentage points (five and a half at 80 percent power), so the contrasts that fail the correction rule out large gaps but not gaps of a point or two.

The geography of the Holm-surviving contrasts is southern and central Europe plus Latin America.

[5] The joint wild-bootstrap Wald statistic on all twenty restrictions is $F(20, 12) = 20.56$. The asymptotic cluster-robust test can impose at most twelve of the twenty restrictions with thirteen clusters and gives $F(12, 12) = 34.26$. Both reject at the one percent level.

Figure 2: Country effects on the recommended equity share



Notes: Point estimates of the country fixed effects from equation (1) (Question 1, Version B, age 40), in percentage points relative to the omitted U.S. baseline. Bars are 95 percent wild cluster bootstrap confidence intervals (Webb six-point weights, 9,999 replications, clustered by model, $G = 13$), the intervals of record, matching Appendix Table A7.

Under the Holm correction, the pre-registered family criterion, thirteen of the twenty contrasts fail to reject equality with the U.S. baseline. Six of those are individually significant on the uncorrected wild-bootstrap test but do not survive the step-down (China, Switzerland, Germany, Turkey, India, and Canada, ranging from -4.4 to -1.1 points), and the remaining seven do not reject even before correction: Australia, Finland, Japan, South Africa, South Korea, Sweden, and the United Kingdom. Among these last seven the four with positive point estimates (Australia, Japan, South Korea, and the United Kingdom) are small and statistically indistinguishable from zero, while Finland and South Africa carry small negative gaps that fall short of significance. The United Kingdom is also the design's pre-registered test of country naming net of language: its contrast is $+0.7$ points with a 95 percent confidence interval of $[-0.8, +2.2]$, inside the pre-registered ± 5 point equivalence margin, so the stricter 90 percent two-one-sided-test interval falls inside it as well and the UK–US equivalence holds. Every contrast that survives the correction is negative.The largest gap against the U.S. baseline level, Brazil's eleven percentage points, moves \$5,500 of the \$50,000 stake from stocks to bonds, and the seven significant contrasts average about six points. For comparison, the gender gap that Foltyn and Olsson (2026) document on the same allocation question is under two percentage points, so country context moves LLM advice by several times the demographic gaps documented so far.

4.3 Language and local framing

I also conduct two within-country tests to disentangle the language from the country framing and to examine if local context shapes the advice. The local-financial-context regression compares Versions C and B across the twenty non-U.S. countries at age 40,

$$y_{imc} \;=\; \alpha_m + \gamma_c + \phi\,\mathbf{1}[\text{version} = C] + \varepsilon_{imc}, \tag{2}$$

and the language regression compares Versions B and B-cross across the fifteen non-English countries,

$$y_{imc} \;=\; \alpha_m + \gamma_c + \lambda\,\mathbf{1}[\text{language} = \text{local}] + \varepsilon_{imc}. \tag{3}$$

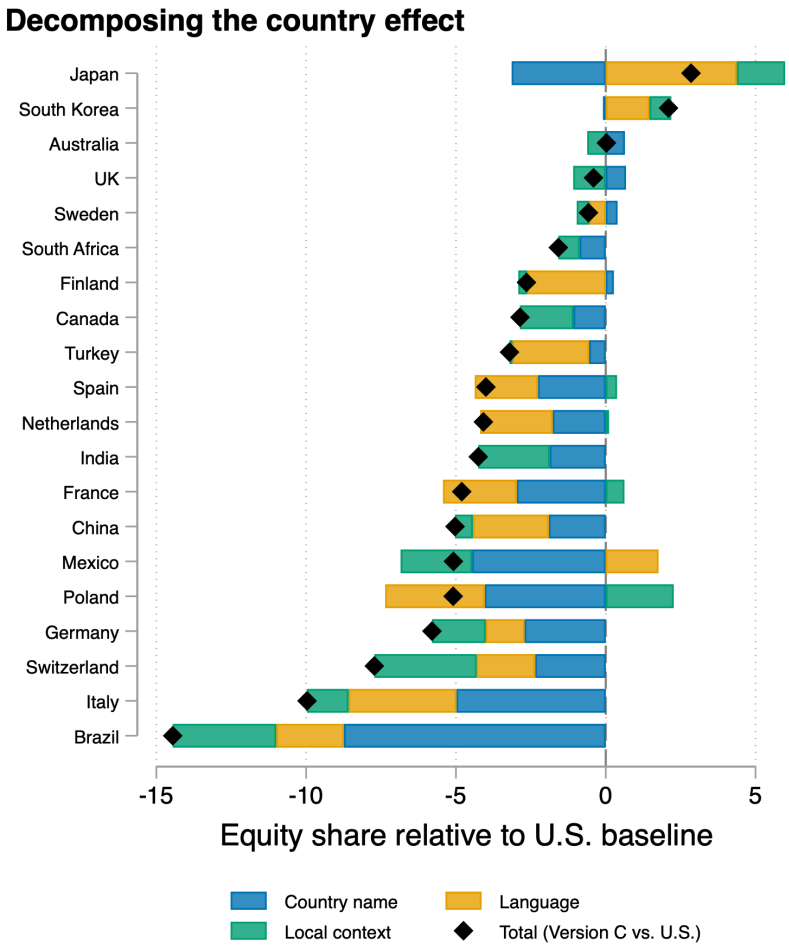
Asking in the local language rather than in English lowers the recommended equity share by 1.3 percentage points ($\hat{\lambda} = -0.013$, wild bootstrap $p = 0.007$, confidence interval -2.3 to -0.4 points). The language effect is statistically significant but economically small, an order of magnitude below the largest country effects. Naming the local instruments does even less. Replacing the generic vocabulary with local pension names, local government bonds, and local index funds moves the allocation by -0.7 points ($p = 0.069$, confidence interval -1.5 to $+0.1$), short of significance at the five percent level.

Together the two estimates decompose the country effect. Language contributes about a point, local framing less than one, so the bulk of every large country effect attaches to the country name itself, which is present in every version. Both small effects are also negative, meaning that each added localization cue nudges the advice toward bonds.

Figure 3 computes a decomposition for individual countries. For each country, I stack the three cues into the full gap from the U.S. baseline, defined as the fully localized Version C relative to the U.S. English generic cell. The country-name slab is the English-only contrast that names the country while holding language and instruments (note that this slab also changes the currency unit and the nominal amount, which move together with the country reference). The language slab is the local-language step on top of it (Version B minus B-cross), and the local-context slab is the local-instrument step (Version C minus B). The three add up to each country's total by construction. In the large-effect countries (Brazil, Mexico, Poland, France, Italy, Germany) the country-name slab accounts for most of the gap, mirroring the pooled decomposition. The language and local-context slabs are usually small, but the per-country pattern is noisier than the pooled $\hat{\lambda}$ and $\hat{\phi}$ and not uniformly negative — Japan is the clearest exception, where a large positive language slab pushes advice toward stocks and produces the panel's biggest positive total.

We can contextualize these results using insights from the computer science literature. The size of the

Figure 3: Decomposing the country effect into naming, language, and local context



Notes: Each bar decomposes a country's fully localized effect on the recommended equity share, Version C relative to the omitted U.S. English generic baseline (Question 1, age 40), into three additive cues in percentage points. Country name is the English-only contrast Version B-cross minus the U.S. (language and instruments held fixed, though the currency unit and nominal amount change with the country reference), language is Version B minus B-cross (country name held fixed), and local context is Version C minus B (language held fixed). The black diamond marks the total, which equals the sum of the three slabs by construction. The six English-prompt countries have no B-cross cell, so their language slab is zero and country name is Version B minus the U.S. Components are model-adjusted cell contrasts from a single saturated country-by-version regression clustered at the model level, reported as a descriptive decomposition, so the per-country language and local-context slabs are noisier than the pooled estimates $\hat{\lambda}$ and $\hat{\phi}$ in the text.

language effect matches what the multilingual-interpretability literature predicts for this design. Tracing studies find that language-specific processing sits in a thin shell around the input and output layers, that the middle layers solving the task operate in a shared concept space lying closer to English than to any other language, and that the degree of cross-lingual sharing rises with model scale. This implies that the sharing should be strongest in the frontier models of this panel (Wendler et al., 2024; Tang et al., 2024; Wu et al., 2025; Lindsey et al., 2025). Under such hub-style processing two translations of the same question become nearly the same query by the middle of the network, so the prediction for equation (3) is a λ near zero, with the small residual consistent with the language-specific features that remain. A

substantial λ was a live possibility ex ante: the same model often asserts different facts depending on the query language (Kassner, Dufter and Schütze, 2021; Qi, Fernández and Bisazza, 2023), so advice retrieved like language-keyed knowledge (German queries drawing on the bond-heavy conventions of German-language financial text) would have inherited that inconsistency. The advice behaves like one policy applied across query languages rather than like language-segmented retrieval.

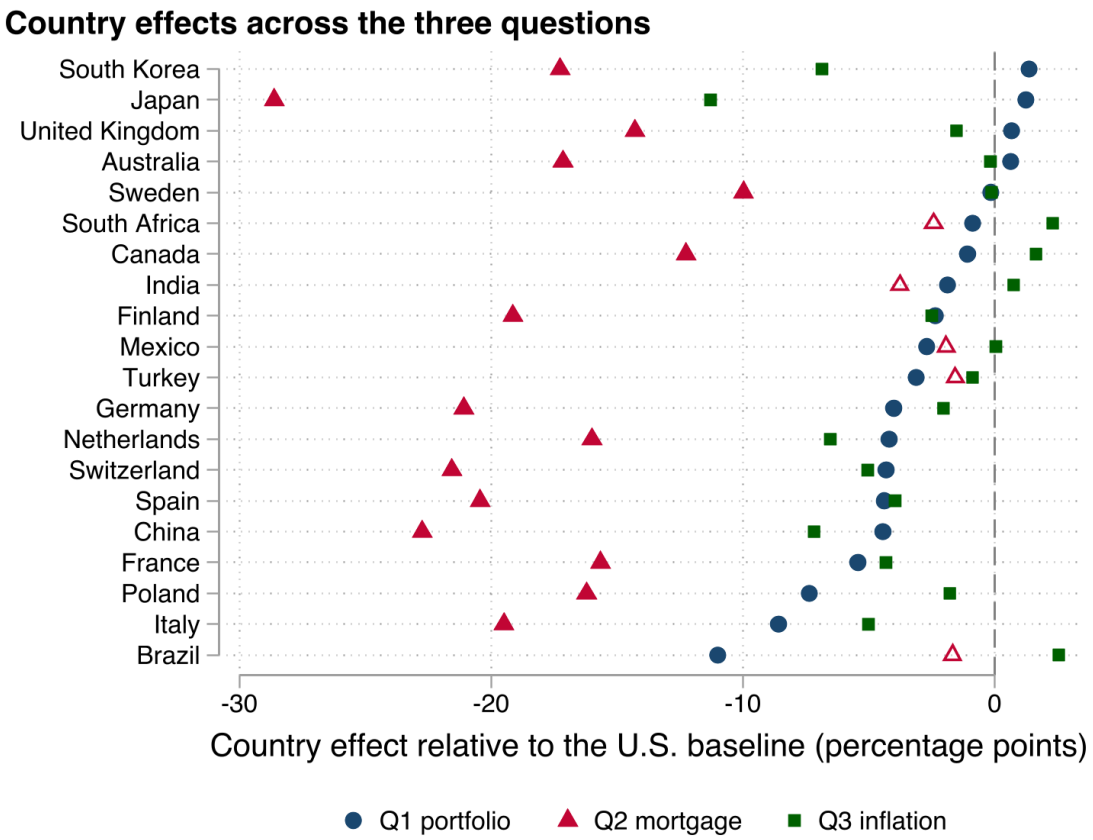
4.4 The mortgage and the inflation questions

I now examine the mortgage and inflation questions. These questions potentially require more local interpretation from the models than the previous allocation problem, in that they require the models to consider mortgage market design and what (potentially local) products to use to hedge against inflation. Appendix Tables A8 and A9 report every contrast with its bootstrap and Holm p-values.

Both questions receive more aggressive advice than the baseline allocation question does. Pooled across the panel, the models put 70 percent of the savings into the index fund rather than into mortgage prepayment and move 74 percent of the deposit into investments against inflation. Compare this to the mean equity share of 54 percent on Question 1. On the mortgage question, all fifteen countries outside the premise-acceptance group receive a recommended investment share that is significantly lower than in the United States, by 10 to 29 percentage points. The five premise-acceptance countries (Brazil, India, Mexico, Turkey, and South Africa), whose prevailing mortgage rates exceeded the prompted 4 percent at the filing date, are the exception. Their contrasts are mostly null, with South Africa (-2.4 points) the only one to survive its family's correction.

On the inflation-protection question, recall that this question came in two parts. The first part asked for a share to invest in an asset, and the second part asked for a recommended instrument to protect against inflation. For the first part, seven countries survive the Holm correction: Japan (-11.3 points), China (-7.2), South Korea (-6.9), the Netherlands (-6.5), Italy (-5.0), Switzerland (-5.0), and France (-4.3). Appendix Table A4 documents the answer to the second part, the primary instrument the model recommends to protect against inflation. Equities are the near-universal primary instrument across most of the panel, recommended in 93 percent or more of responses in fourteen countries and a clear majority in all but the highest-inflation economies. In Brazil, Turkey, and Mexico the models switch to inflation protection. Brazil's recommendations are dominated by inflation-linked bonds (86.7 percent), Turkey's by inflation-linked bonds and gold (66.0 and 16.3 percent, 82.3 percent combined), and Mexico's tilt the same way (58.1 percent inflation-linked bonds), with Poland split (38.6 percent inflation-linked). The tilt also appears inside the English-prompting group, where models in the two higher-inflation members, India and South Africa, recommend inflation-protection (11.5 and 16.3 percent, compared to under 3 percent in the United States, the United Kingdom, Canada, and Australia).

Figure 4: Country effects across the three questions



Notes: Estimated country fixed effects on all three questions (Version B, age 40), in percentage points relative to the omitted U.S. baseline, for the twenty non-U.S. countries. Countries are ordered by the Question 1 effect, matching Figure 2. Hollow triangles mark the five premise-acceptance countries, whose prevailing mortgage rate exceeded twice the prompted 4 percent, so their mortgage contrasts test premise acceptance rather than calibration. Confidence intervals are omitted here because sixty intervals on one axis are unreadable. The Question 1 intervals are plotted in Figure 2, and all three sets of estimates are tabulated with their confidence intervals in Appendix Tables A7, A8, and A9. The registration's Q4 is the manuscript's Question 3.

Figure 4 compares the answer to each of the three questions for each country. Countries are ordered by the answer to the first portfolio question, as in Figure 2. The first takeaway from the figure is that the ordering of the country effect on Question 1 does not carry through to the other questions. It is also noteworthy that the mortgage question provides a much larger country effect, with every one of the fifteen countries sitting at least ten points below the U.S. baseline. The cross-country correlation of the estimated effects is 0.09 between the portfolio and mortgage questions and 0.04 between the portfolio and inflation questions across the fifteen calibration countries. The two secondary questions do correlate strongly with each other, however. The country effects correlate at 0.76 across the same fifteen countries.[6]

An interesting example is Japan. On the first portfolio question, Japan is very close to the United States (in fact above the United States, although not at statistically significant levels). However, Japan is

[6]These correlations are descriptive, computed across countries on the estimated fixed effects, and are not pre-registered.

the most negative country on both the mortgage question and the inflation question. A fixed nominal mortgage is a different asset under Japanese inflation history. In general, the wider range of outcomes under the secondary questions suggest that models reach for country-specific economic content when it is necessary. However, almost none of that content reaches the first portfolio question.

4.5 Age gradients and pension generosity

Question 1 varies the investor's age across 30, 40, and 55 in Version B. The age regression adds age effects and the pre-specified interaction with pension replacement rates:

$$y_{imca} \;=\; \alpha_m + \gamma_c + \delta_a + \psi\,\tilde{\rho}_c\,\mathbf{1}[\text{age} = 55] + \varepsilon_{imca}, \tag{4}$$

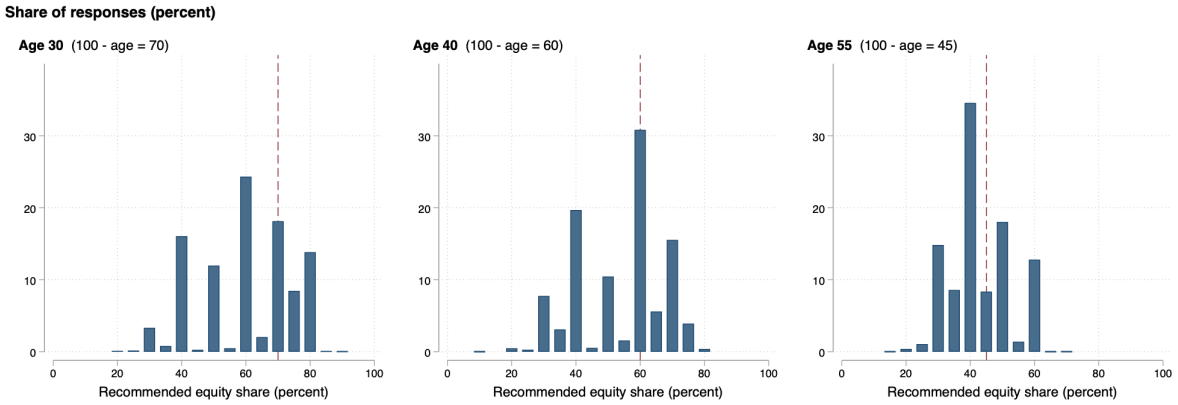
where $\tilde{\rho}_c$ is the OECD net pension replacement rate standardized across the twenty-one countries. The test asks whether the age-55 tilt toward bonds is attenuated where pensions replace more income, as background-wealth logic requires: a generous pension is a large bond-like claim whose weight in total wealth grows as retirement nears, so the equity share of the liquid portfolio should fall less steeply with age in high-replacement countries.

The age gradient itself is strong and conventional. Relative to the 40-year-old, the 30-year-old is recommended 6.24 percentage points more equity and the 55-year-old 11.13 points less. Its interaction with pension generosity is an exact zero. The estimate of ψ is -0.0002 ($p = 0.96$), and the coarser above-median replacement-rate split gives $+0.4$ points ($p = 0.49$). Appendix Table A10 reports the full regression alongside a specification without the pension interaction. The models apply the same glide path in the panel's most generous pension systems and in its least generous, and Appendix Figure A1 shows it applied nearly verbatim in all twenty-one country panels.

We can further think about rules-of-thumb in financial advice from LLMs by asking how literally the advice follows a standard rule-of-thumb. Figure 5 plots the distribution of the recommended share at each prompted age against the 100-minus-age prescription of 70, 60, and 45 percent equity. The rule describes the pooled advice exactly at one age. At 40 the modal recommendation coincides with the rule's 60 percent.

At the other two ages the mode is different from the rule, but tends to congregate at the nearest round number. The 30-year-old's most common recommendation is still 60 percent (24 percent of responses) rather than the rule's 70 (18 percent), and the 55-year-old's is 40 percent (35 percent of responses) rather than the rule's 45, which receives only 8 percent. The slopes tell the same story. Pooled advice falls by 0.70 percentage points per year of age against the rule's exact 1.00, nearly identically in the English-prompt countries (-0.65) and elsewhere (-0.72), and all thirteen models tilt downward in age

Figure 5: The advice against the 100-minus-age rule, by prompted age



Notes: Distribution of the recommended equity share at each prompted age (Question 1, Version B, all twenty-one countries and thirteen models, parsed non-refusals, $N = 40{,}910$). Dashed lines mark the 100-minus-age prescription. Essentially all responses sit on multiples of five percentage points, so each panel is a discrete bar chart as in Figure 1. Exploratory exhibit, not pre-registered.

with ten of the thirteen flatter than the rule (model-level slopes span -0.23 to -1.33). The glide path is thus not literal 100-minus-age compliance, although it is relatively close to the rule.

The deviations from the rule also have a normative direction. Love (2013) shows that the 100-minus-age prescription tracks the optimal decline in equity exposure through the middle of the life cycle but errs at both tails, where the optimum lies above the rule: the young would hold more than 100 percent equity absent the borrowing constraint, and the optimal share flattens or even rises late in life as background risk recedes. The models anchor at age 40, where the rule is most defensible, and deviate from it toward conservatism at both ends, so where the advice departs from the heuristic it departs on the opposite side from the life-cycle optimum. This is the same early-life conservatism that Choukhmane et al. (2026) find a calibrated life-cycle model cannot rationalize.

The explanation text adds a telling asymmetry: explicit citations of the heuristic ("rule of thumb," "100 minus age") appear in 15 to 17 percent of primary-cell explanations in the United States and India, but in under 1 percent of explanations outside the English-prompt countries, even though the allocations themselves bunch identically. This is consistent with a single advice policy formed in the English-language financial-planning corpus and applied everywhere, although other explanations are also plausible.

4.6 Robustness

The pre-registration commits to robustness checks on two fronts: first, a temperature check on the models, and second, a check that examines the impact of the local currency. These are described below. Further, Replacing the thirteen model fixed effects with six vendor fixed effects, for comparability with Foltyn and Olsson (2026), leaves every country contrast essentially unchanged, although with six clusters the variant supports no formal inference.

The first temperature check re-collects the primary cell at the temperature bookends of 0 and 1.0 for five countries chosen to span the range of estimated effects (the United States, Germany, Sweden, Japan, and Brazil), asking whether the results are specific to the pre-registered sampling temperature of 0.7. Appendix Table A11, Panel A shows that they are not. The pooled mean equity share shifts by 0.3 percentage points at either bookend relative to the 0.7 baseline (wild bootstrap $p = 0.55$ at temperature 0 and $p = 0.40$ at 1.0), and the country contrasts barely move. Brazil's contrast is -10.0, -11.0, and -11.4 points at temperatures 0, 0.7, and 1.0, each with $p < 0.001$, Germany's ranges from -3.6 to -4.2, and Japan's and Sweden's are statistically indistinguishable from zero at every temperature. The only thing that temperature affects is the dispersion around the answer, and only mildly, with the mean within-cell standard deviation rising from 0.035 at temperature 0 to 0.041 at 1.0.

The second robustness check addresses a numeral-magnitude confound. Version B states the stake as a round native-currency figure, so the cross-country contrasts compare prompts whose numerals differ by orders of magnitude. For example, the number is 7,500,000 yen in Japan against 50,000 dollars in the baseline US cell. An advisor that anchored on the size of the number rather than on the economics could generate spurious country effects. For Sweden, Japan, and Germany I therefore re-collect Version B with the amount stated as $50,000, holding country and language fixed (Appendix Table A11, Panel B). The pooled shift is a tight zero (-0.1 percentage points, wild bootstrap $p = 0.84$, confidence interval -1.6 to $+1.4$ points). The one significant per-country shift is Japan, whose seven-digit yen figure raises the recommended share by 2.8 points relative to the dollar statement ($p = 0.01$). Japan's country contrast is small and statistically null with native numerals ($+1.2$ points) and remains small under the xdollar statement (an implied -1.6 points). Numeral anchoring cannot manufacture the eleven-point Brazil contrast or the geographic pattern of the Holm survivors.

I run a further series of checks on the analysis-side, but these do not change the results. This is largely because the data leave it nothing to do. The three pre-registered refusal-handling checks — a linear probability model of refusal on country and model effects, a re-estimation of equation (1) with refusals imputed as zero equity, and Lee-style trimming bounds triggered if country-level refusal rates spread by more than two percentage points — are all empty, because the collection produced no refusals in 267,150 responses (Section 3.7). The refusal outcome is constant, the imputation changes no observation and returns the primary estimates exactly, and the trimming trigger is unmet, as is the Q3 refusal contingency.

5 A Normative Benchmark: What Should the Advice Be?

The experimental results establish *whether* LLM advice varies across countries. They cannot, by themselves, say whether that (limited) variation represents appropriate advice. A lower recommended equity

share in Japan than in the United States is a bias if the economically appropriate advice is the same in both countries, and sound calibration if Japanese institutions or preferences genuinely warrant more conservative advice. Separating these interpretations requires a statement of what the advice *should* be in each country. This section provides that statement in the simplest possible model and shows how to take it to the experimental data.

The benchmark delivers three empirical objects. First, a country-specific optimal equity share α_c^* against which the level of LLM advice can be compared. Second, a *calibration slope* that measures how much of the cross-country variation that should appear in the advice actually does appear, nesting the "U.S.-default" hypothesis (H2) as the case where the slope is zero. Third, a welfare metric that converts the gap between actual and optimal advice into a certainty-equivalent return loss, and into dollars on the \$50,000 stake in the prompt. The model is deliberately static and closed form. A full life-cycle model in the tradition of Cocco, Gomes and Maenhout (2005) would add realism on several margins discussed in Section 5.9, but at the cost of a large calibration apparatus (see e.g. Choukhmane et al., 2026). However, the central mechanism (background wealth crowding equity into the liquid portfolio) is captured in a simpler manner here.

5.1 The investor's problem

Consider the investor described in the prompt: age $a = 40$, liquid savings W (the local-currency equivalent of \$50,000), choosing how to split W between a broad equity index fund and domestic government bonds. Let r_c denote the real return on the safe asset in country c, let the equity fund earn an expected real excess return $\mu_c - r_c$ with standard deviation σ_c, both measured in local currency, and let the investor have constant relative risk aversion (CRRA) preferences over terminal wealth with risk-aversion coefficient γ,

$$U(W_T) \;=\; \frac{W_T^{1-\gamma}}{1-\gamma}. \tag{5}$$

The CRRA assumption is the standard one in household finance and has a convenient implication used below: the investor cares about *proportional* gains and losses, so the optimal portfolio *share* does not depend on the level of wealth.

The classic result of Merton (1969) and Samuelson (1969) is that the optimal share of the portfolio held in equity is

$$\theta_c^* \;=\; \frac{\mu_c - r_c}{\gamma\,\sigma_c^2}. \tag{6}$$

Each of these variables map to a country-level parameter that the data can help us discipline. The numerator is the compensation for holding equity risk: a higher local-currency equity premium raises the optimal share linearly, scaled by $1/(\gamma\,\sigma_c^2)$. The denominator is the cost of holding that risk: it rises

with the variance of equity returns σ_c^2 (riskier markets warrant smaller positions) and with risk aversion γ (more risk-averse investors hold less equity). This is a convenient benchmark because of the result that the equity share is independent of the investment amount. In addition, if we assume that returns are independent over time, the equity share is also indepedent of the horizon. This later result is convenient here: the prompt's five-year horizon does not require a separate model input. Time-varying expected returns would break this irrelevance, a point I return to in Section 5.9.

Because the prompt restricts the answer to a fraction between zero and one, the benchmark is capped to the same interval,

$$\alpha_c^* \;=\; \min\{1,\; \max\{0,\; \theta_c^*\}\}, \tag{7}$$

which matters in practice for the integrated benchmark below.

5.2 Benchmark S: the stake as a standalone portfolio

The first benchmark takes the prompt at face value: the investor is asking about this $50,000 as a self-contained decision (the investor is 'looking to invest $50,000 over a five-year horizon."). Under Benchmark S the optimal share is simply equation (6),

$$\alpha_c^S \;=\; \left[\frac{\mu_c - r_c}{\gamma_c\, \sigma_c^2}\right]_0^1, \tag{8}$$

where $[x]_0^1 \equiv \min\{1, \max\{0, x\}\}$ denotes truncation to the unit interval. Cross-country variation in α_c^S comes from two sources: local-currency equity volatility and, when preferences are allowed to vary, risk aversion. Under the world-index reading developed below, the expected excess return is common across currencies by uncovered interest parity, so it sets the level of every benchmark but not the spread across them. At the baseline $\bar{\gamma} = 3$, equation (8) produces interior optimal shares from about 0.21 in Turkey to 0.53 in South Korea, a band that overlaps the 0.46 to 0.58 range in which LLM recommendations fall, so deviations are informative in both directions. Choukhmane et al. (2026) estimate by simulated method of moments a coefficient of relative risk aversion of about 5 (5.2 under their survey prompts, 4.7 under the academic prompt) that rationalizes the advice of current-generation models, which corresponds to the $\gamma = 5$ point of the sensitivity grid, one step above the $\bar{\gamma} = 3$ baseline. At $\gamma = 2$ the U.S. benchmark rises to about 0.78, and at $\gamma = 10$ shares compress to between 6 and 16 percent across countries — below 8 percent for Turkey and Brazil — implausibly low relative to observed advice, which is itself useful information when reading the sensitivity tables.

How can we measure μ_c and σ_c in this context? The prompt says "a broad stock market index fund," which could mean the local equity market or a global fund held from country c. Instead, I take world equity index measured in local currency as the baseline. This is a global representation of the market

return, and because it is measured in local currency, the cross-country variation in α_c^S enters through local-currency volatility rather than through the equity market itself. This matches the instruments named in Version C (MSCI World trackers in most countries) and avoids attributing "bias" to any sensible reluctance to recommend concentrated exposure to a small home market.

With a world index and a common γ, essentially all cross-country variation in α_c^S comes from differences in local-currency volatility, that is, from currency covariances and the hedged-versus-unhedged choice. The baseline uses unhedged local-currency moments, and a currency-hedged variant accompanies it as a robustness row. The equity premium is held common across countries at an excess return of about 5.3 percent, the arithmetic mean of the MSCI World index (net, in U.S. dollars, 1999–2025) over the U.S. long-term rate. For the U.S. cell the local-currency moments are an equity volatility near 18 percent and a real safe rate near 0.4 percent (trailing twenty-year mean), taken from MSCI and the Global Macro Database (Appendix Table A12); at the baseline $\bar{\gamma} = 3$ they imply an interior Merton share of about 0.52 and provide a calibration cross-check on α_{US}^S.

5.3 Benchmark I: adding pension wealth and human capital

The prompt also says "I have a standard pension plan otherwise," which invites the model to consider the investor's full balance sheet. The second benchmark therefore integrates two non-tradable assets that a 40-year-old holds outside the liquid portfolio: pension wealth P_c, the present value of future pension benefits, and human capital H_c, the present value of future labor income (Bodie, Merton and Samuelson, 1992; Campbell and Viceira, 2002; Gálvez and Paz-Pardo, 2026).

Specifically, the investor's total wealth is $W + H_c + P_c$, and by equation (6) she wants a fraction θ_c^* of *total* wealth exposed to equity. If pension claims are bond-like (a reasonable description of public and defined-benefit pillars) and a fraction ζ of human capital is equity-like, the non-tradable assets already contribute ζH_c of equity exposure, and the entire remaining equity demand must be met inside the liquid account. The optimal share of the *liquid* portfolio is then

$$\alpha_c^I \;=\; \min\Big\{1,\; \max\Big\{0,\; \underbrace{\theta_c^* \frac{W + H_c + P_c}{W}}_{\text{scale up: outside wealth is safe}} \;-\; \underbrace{\zeta \frac{H_c}{W}}_{\text{scale down: wage risk is equity-like}} \Big\}\Big\}. \tag{9}$$

The first term says that safe outside wealth makes the investor effectively less exposed to her liquid portfolio, so she should hold *more* equity in it. The more generous the pension system, the stronger this push. The second term works in the opposite direction when labor income co-moves with the stock market. In this setup, the equity-like part of human capital is treated as perfectly correlated with the market and human capital is discounted at the safe rate. Finally, treating pension claims as bond-like is itself contestable for pay-as-you-go systems whose benefits are indexed to wages.

The two balance-sheet inputs, pension wealth P_c and human capital H_c, have standard annuity expressions in the net replacement rate, the average wage, the retirement age, life expectancy, and real wage growth, with pension wealth discounted from retirement and human capital summed over the remaining working years. Appendix B.5 states the two formulas (equations (15) and (16)) and their calibration assumptions. As the persona in the prompt reveals no wage, the average earner is the natural calibration target.

Benchmark I delivers a sharp and useful prediction: because $H_c + P_c$ for an average 40-year-old is an order of magnitude larger than W, the multiplier in equation (9) pushes α_c^I to the corner of 100 percent equity in essentially every country. Under the calibration the corner is robust across the baseline region of the sensitivity grid: at $\zeta = 0$ it holds for all 21 countries through $\bar{\gamma} = 5$ and breaks at $\bar{\gamma} = 10$ only for four countries (Brazil, India, Mexico, South Africa). At $\zeta = 0.2$ it likewise holds everywhere through $\bar{\gamma} = 3$.

This benchmark has two important implications. First, to the extent LLMs reason in an integrated, full-balance-sheet way, their advice should be near 100 percent everywhere, and observed cross-country *variation* cannot be rationalized by Benchmark I at all. The integrated model converts any cross-country variation into prima facie evidence against normative calibration. Second, advice below α_c^S is conservative under the standalone reading and the baseline integrated calibration, and advice between α_c^S and one is rationalizable only with integrated reasoning.

Making earnings substantially riskier, as in Gálvez and Paz-Pardo (2026), does generate cross-country variation in α_c^I. At an equity loading of roughly half of human capital ($\zeta \in \{0.4, 0.6\}$), the low-multiplier countries leave the corner first, Brazil and Mexico at $\zeta = 0.4$, joined by Poland, South Africa, Japan, and India at $\zeta = 0.6$ under the baseline $\bar{\gamma} = 3$. However, this result is fragile, and I do not treat it as the benchmark for two reasons. First, exit from the corner is knife-edge, because equation (9) differences two terms that are each an order of magnitude larger than the unit interval. A country passes from the corner at a 100% equity share through the interior to zero within a step or two of the grid (Brazil is an example of this).

Second, the ζ penalty attaches only to H_c, so a country whose outside wealth is dominated by pension promises stays at 100 percent equity no matter how risky earnings are made. Turkey is the most extreme example of this, with the high-wage-growth and low-real-rate calibration its recent macro history produces. The extended ζ grid therefore stays in the released code and benchmark output for transparency, but the confrontation with the data in section 5.5 runs on α_c^S, and Benchmark I enters only through the bracketing logic above.

5.4 Preferences: common versus locally calibrated risk aversion

I compute every benchmark under two preference regimes. Under the *common-preference* regime, $\gamma_c = \bar{\gamma}$ for all countries, with $\bar{\gamma} \in \{2, 3, 5, 10\}$ as the sensitivity grid. Cross-country variation in the benchmark then reflects institutions and asset markets only. Under the *preference-calibrated* regime, risk aversion varies with the country's measured willingness to take risks from the Global Preferences Survey (Falk et al., 2018). We map the GPS risk-taking index into risk aversion as

$$\gamma_c \;=\; \bar{\gamma}\,\exp\!\big(-\eta\,(z_c - z_{US})\big), \tag{10}$$

where z_c is the GPS risk-taking score, which is standardized at the individual level and aggregated to the country mean by Falk et al. (2018). $\eta \geq 0$ governs how strongly measured preferences translate into the benchmark. Concretely, η is the semi-elasticity of risk aversion with respect to the standardized GPS score ($d \ln \gamma_c / dz_c = -\eta$), so a one-standard-deviation increase in measured willingness to take risks scales risk aversion by the constant factor $e^{-\eta}$ wherever the country sits. The functional form keeps γ_c positive and makes the U.S. benchmark invariant to η.

The mapping is admittedly ad hoc, as there is no consensus conversion from survey risk measures to CRRA coefficients. I therefore treat η as a transparent sensitivity parameter on a grid such as $\{0, 0.3, 0.6\}$, where $\eta = 0$ recovers the common-preference regime exactly. At $\eta = 0.6$ the implied U.S.–Japan risk-aversion ratio is roughly 1.3, a moderate magnitude relative to documented preference heterogeneity. The end result is that we can see which benchmark the LLM advice tracks: if the advice tracks the preference-calibrated benchmark but not the common-preference one, models are responding to cultural risk attitudes.

5.5 Taking the benchmark to the data

The benchmark is computed from external country-level inputs and then confronted with the experimental estimates using three measures: the advice gap, the calibration slope and the welfare cost of the gaps. Appendix Table A12 reports every country-specific input the benchmark requires, with the common parameters and sources in the table notes. It lists the eight country-varying portfolio inputs, the equity premium being common across countries, the two mortgage-question inputs of deductibility and expected inflation, and the resulting standalone benchmark share α_c^S. Three common parameters complete the calibration: baseline risk aversion $\bar{\gamma}$ (grid $\{2, 3, 5, 10\}$), the preference-mapping intensity η (grid $\{0, 0.3, 0.6\}$), and the equity loading of human capital ζ (baseline 0, robustness 0.2).

Advice gaps. For each country, the gap between mean LLM advice and the benchmark,

$$\Delta_c \;=\; \bar{y}_c - \alpha_c^S, \tag{11}$$

where $\bar{y}_c$ is the mean recommended fraction in the primary cell (Version B, Q1, age 40). The gap Δ_c is a descriptive level comparison. The wild-cluster-bootstrap confidence interval inherited from the country fixed effects describes the advice deviation from the U.S. baseline, not the level gap itself. Note that LLM advice is known to anchor on round-number saving rates and rules of thumb such as the 4 percent withdrawal rule (Choukhmane et al., 2026), heuristics that shift $\bar{y}_c$ away from any optimizing benchmark by a roughly common amount across countries. Such a common shift enters Δ_c in every cell but largely differences out of the calibration slope in equation (12) below, which is identified off cross-country movement rather than levels.

The calibration slope. The next measure examine whether cross-country *movements* in the benchmark show up in the advice. Let $\hat{b}_c$ denote the estimated country fixed effect from equation (1) (the LLM advice deviation from the U.S. baseline, written γ_c there but renamed here to avoid collision with risk aversion) and let $\delta_c^* = \alpha_c^S - \alpha_{US}^S$ denote the benchmark's prescribed deviation. The regression

$$\hat{b}_c \;=\; \kappa_0 + \kappa\, \delta_c^* + e_c \tag{12}$$

summarizes calibration in a single number. Full normative calibration implies $\kappa = 1$: advice moves one for one with what local fundamentals prescribe. The U.S.-default hypothesis implies $\kappa = 0$. Note that the converse does not hold, since large country effects orthogonal to fundamentals also produce a zero slope. I therefore report the gap table and the slope together. Intermediate values measure partial calibration.

Welfare cost of the gaps. CRRA preferences yield a closed-form price tag for deviating from the optimum. The certainty-equivalent return of holding share α is quadratic, $ce(\alpha) = r_c + \alpha(\mu_c - r_c) - \frac{1}{2}\gamma_c\alpha^2\sigma_c^2$, so the annual loss from holding α instead of the benchmark is computed directly as

$$\mathcal{L}_c \;=\; ce(\alpha_c^*) - ce(\alpha) \;=\; \tfrac{1}{2}\,\gamma_c\,\sigma_c^2\left[\left(\theta_c^* - \alpha\right)^2 - \left(\theta_c^* - \alpha_c^*\right)^2\right], \tag{13}$$

which reduces to the familiar $\frac{1}{2}\gamma_c\sigma_c^2(\alpha-\alpha_c^*)^2$ whenever the benchmark is interior ($\alpha_c^* = \theta_c^*$). This formula remains correct when the cap binds, as it does throughout Benchmark I, where the simpler expression would understate losses materially.

Two features are worth emphasizing when reading the results. The loss is locally quadratic, so small

advice gaps are nearly free while large ones are disproportionately costly. Around an interior optimum it is symmetric, so excessive conservatism and excessive aggression are priced on the same scale. Evaluated at the prompt's stake, the dollar cost over the five-year horizon is approximately $W\left[(1+\mathcal{L}_c)^5-1\right] \approx 5\,W\mathcal{L}_c$. At plausible calibrations a 10 percentage point advice gap costs single-digit basis points per year, on the order of $100 over five years on $50,000. This is still a relatively small number, and hence the headline welfare object is the cross-country *incidence* of losses rather than their absolute size.

The life-cycle literature puts these numbers in context. Simple allocation rules are cheap for the textbook household because the value function is flat near the optimum — Love (2013) puts the annual loss from the best fixed rule below $200, and Bagliano, Fugazza and Nicodano (2021) and Khemka, Steffensen and Warren (2021) find losses of a fraction of annual consumption for standard prescriptions. The losses become first-order under the conditions this paper documents, however: applying a path designed for one risk-aversion level to an investor with another level of risk aversion (Khemka et al., 2021), and a labor-income risk that departs from the textbook process, raises losses to 3 to 9 percent of annual consumption (Bagliano et al., 2021). The single-digit-basis-point figure above therefore describes the benign textbook case rather than an upper bound.

5.6 Advice gaps and the calibration slope

The benchmark prescribes large differences in advice across this panel. At the baseline grid point ($\bar{\gamma} = 3$, $\eta = 0$), the standalone benchmark α_c^S ranges from 0.21 in Turkey and 0.23 in Brazil, whose local-currency volatilities are the highest in the panel, to 0.53 in South Korea and 0.52 in the United States, a prescribed spread of thirty-two percentage points. Mean advice spans twelve points, from 0.46 to 0.58 (Table 1), so the advice covers roughly a third of the spread that local fundamentals warrant. In addition, Turkey, whose prescribed share is the lowest in the panel, receives mean advice of 0.54, statistically indistinguishable from the U.S. baseline and near the top of the realized band.

Table 2 reports the per-country results. The gaps are positive nearly everywhere, and they are largest where fundamentals depart most from the U.S. baseline: 33 percentage points in Turkey, 23 in Brazil, 17 in Mexico, and 15 in Japan and Poland. Across most of western Europe, whose prescribed shares sit close to the U.S. value, the gaps are within a few points of zero, with France exact to within a fifth of a point.

Figure 6 plots the estimated country effects against the benchmark's prescribed deviations and fits the calibration regression (12). The slope is $\hat{\kappa} = 0.130$ (standard error 0.112). The primary wild bootstrap over countries rejects full calibration decisively ($p < 0.001$ against $\kappa = 1$) and does not reject the U.S.-default anchor ($p = 0.317$ against $\kappa = 0$, 95 percent confidence interval $[-0.145, 0.328]$). Taken at face value, the advice transmits about an eighth of the cross-country variation that local fundamentals

Table 2: Advice gaps against the standalone benchmark and their welfare cost

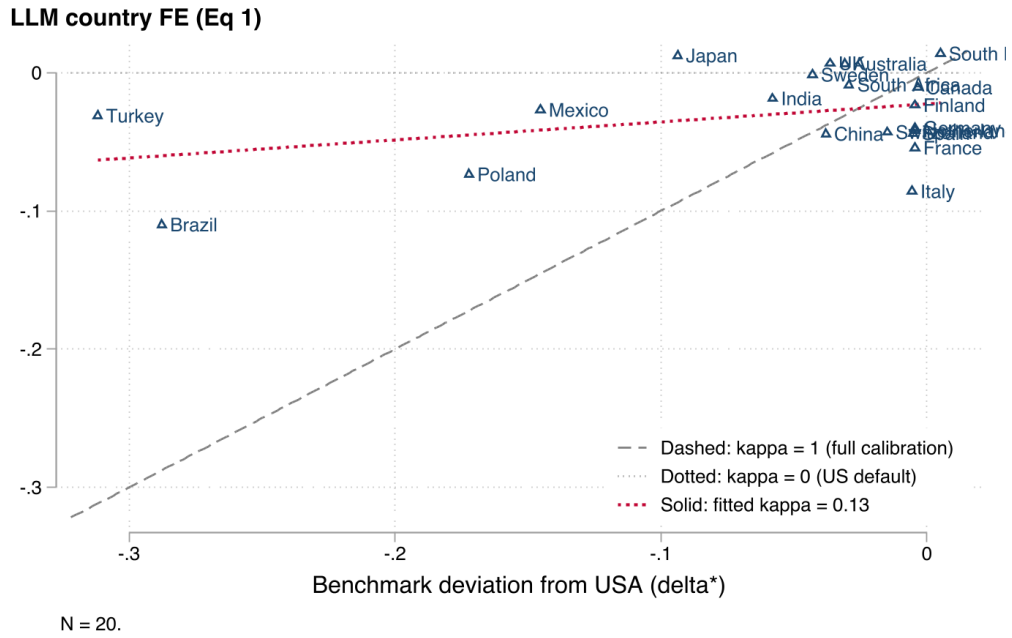
Country	Mean advice	Benchmark α_c^S	Gap Δ_c	Loss (bp/yr)	\$ over 5y
Turkey	0.54	0.21	0.33	136.7	3,513
Brazil	0.46	0.23	0.23	57.8	1,461
Mexico	0.54	0.38	0.17	19.5	489
Poland	0.50	0.35	0.15	16.3	408
Japan	0.58	0.43	0.15	14.7	368
United Kingdom	0.58	0.49	0.09	4.5	113
Sweden	0.57	0.48	0.09	4.4	110
India	0.55	0.46	0.09	4.3	109
Australia	0.58	0.49	0.08	3.9	98
South Africa	0.56	0.49	0.07	2.5	63
South Korea	0.58	0.53	0.06	1.6	40
USA	0.57	0.52	0.05	1.2	29
China	0.52	0.48	0.04	0.9	23
Canada	0.56	0.52	0.04	0.8	21
Italy	0.48	0.52	−0.03	0.5	14
Finland	0.55	0.52	0.03	0.4	11
Switzerland	0.53	0.51	0.02	0.2	5
Germany	0.53	0.52	0.01	0.1	2
Netherlands	0.53	0.52	0.01	0.1	1
Spain	0.53	0.52	0.01	0.0	1
France	0.52	0.52	−0.00	0.0	0

Notes: Mean recommended equity share by country in the primary cell (Question 1, Version B, age 40) against the standalone benchmark α_c^S at the baseline grid point ($\bar{\gamma} = 3$, $\eta = 0$, world index in local currency). The gap is $\Delta_c = \bar{y}_c - \alpha_c^S$. The annual loss prices the gap with the certainty-equivalent formula (13) at the same grid point, in basis points of the liquid stake per year, and the last column converts it to dollars over the prompt's five-year horizon on the \$50,000-equivalent stake. Countries are ordered by annual loss.

prescribe.

The slope's distance from full calibration also survives treating the benchmark regressor as measured with error, the concern that historical parameter uncertainty in δ_c^* attenuates $\hat{\kappa}$ toward zero. Because δ_c^* is a deterministic function of the local-currency return volatility σ_c, estimated from the same twenty-seven annual return years (1999–2025), classical measurement error is the right frame. Attenuation biases the forward regression down, so the structural slope is bracketed between $\hat{\kappa} = 0.13$ and the reciprocal of the reverse regression of δ_c^* on the country effects. That Frisch bound is $[0.13, 0.97]$: even its upper end, which attributes the entire off-line scatter of the country effects to mismeasured fundamentals, stays below the unit slope of full calibration. The bracket is wide because the advice barely moves. The estimated country effects span twelve percentage points against the thirty-two the benchmark prescribes ($R^2 \approx 0.14$), so the distance from full calibration is a fact about the advice rather than an artifact of a precisely measured regressor. A complementary check isolates the volatility-sampling channel directly: resampling the twenty-seven return years with replacement and recomputing every δ_c^* on each draw, holding the year set common across countries so the shared world-index return cancels in the difference, shifts the slope only from 0.13 to a resampled center of 0.10. That exercise prices in volatility sampling error alone. Uncertainty in the common equity premium enters δ_c^* through the ratio $(\mu - r)/\gamma$ and so

Figure 6: The calibration slope: advice deviations against prescribed deviations



Notes: Estimated country fixed effects $\hat{b}_c$ from equation (1) plotted against the benchmark's prescribed deviations $\delta_c^* = \alpha_c^S - \alpha_{US}^S$ at the baseline grid point ($\bar{\gamma} = 3$, $\eta = 0$), for the twenty non-U.S. countries. The dashed 45-degree line is full calibration ($\kappa = 1$), the dotted horizontal line is the U.S.-default anchor ($\kappa = 0$), and the solid line is the fitted regression (12) with slope $\hat{\kappa} = 0.13$. Inference of record is the wild bootstrap over countries reported in the text.

maps onto the common risk-aversion grid of the next paragraph, where full calibration is rejected at every grid point whose prescribed levels overlap the realized advice.

The slope is also not an artifact of the baseline grid point. Across the full $(\bar{\gamma}, \eta)$ grid, $\hat{\kappa}$ ranges from 0.07 to 0.43 and is never near one. The rise with $\bar{\gamma}$ is mechanical, since higher common risk aversion compresses the prescribed deviations that form the regressor. At $\bar{\gamma} = 10$ the point estimate reaches 0.43 with an interval too wide to reject full calibration, but that grid point prescribes equity shares below 16 percent in every country (as low as 6 percent in Turkey and Brazil), implausibly far below any advice observed in the data (Section 5.2). At every grid point whose prescribed levels overlap the realized advice, $\hat{\kappa}$ is at most 0.22 and full calibration is rejected with $p < 0.001$. Allowing risk aversion to vary with measured preferences does not raise the slope either: at $\bar{\gamma} = 3$ the preference-calibrated values are 0.12 at $\eta = 0.3$ and 0.09 at $\eta = 0.6$, slightly below the common-preference estimate.

5.7 The welfare cost of the country gaps

For the median country the gap costs 1.6 basis points of the liquid stake per year, about $40 over the prompt's five-year horizon on the $50,000-equivalent stake, and sixteen of the twenty-one countries lie below five basis points per year (Table 2). The loss is locally quadratic, so the moderate gaps that dominate the panel are nearly free. This is the sense in which the compression finding, taken on its own,

is individually cheap.

However, the incidence of these losses differs widely. The countries where the gap is expensive are exactly those whose fundamentals sit farthest from the U.S. baseline that anchors the advice. Turkey pays 137 basis points per year, about $3,500 over five years or seven percent of the stake, Brazil pays 58 basis points, Mexico 19, Poland 16, and Japan 15, while no western European country pays more than five. This ordering is what a single, baseline-anchored rule predicts. When one policy anchored near the U.S. baseline is applied everywhere, the deviation it leaves uncorrected grows with the distance between local fundamentals and that baseline, and because the loss is quadratic in that deviation the cost falls not on the average user but on the households whose circumstances least resemble the environment in which the rule was formed. Because AI advice is concentrated in a handful of foundation models trained on the same U.S.-English corpus, the misallocation is correlated rather than diversified across the households exposed to it. Of course, these figures assume the user adopts the advice, that the standalone benchmark is the right target, and that the allocation is static, so they price the advice rather than realized behavior.

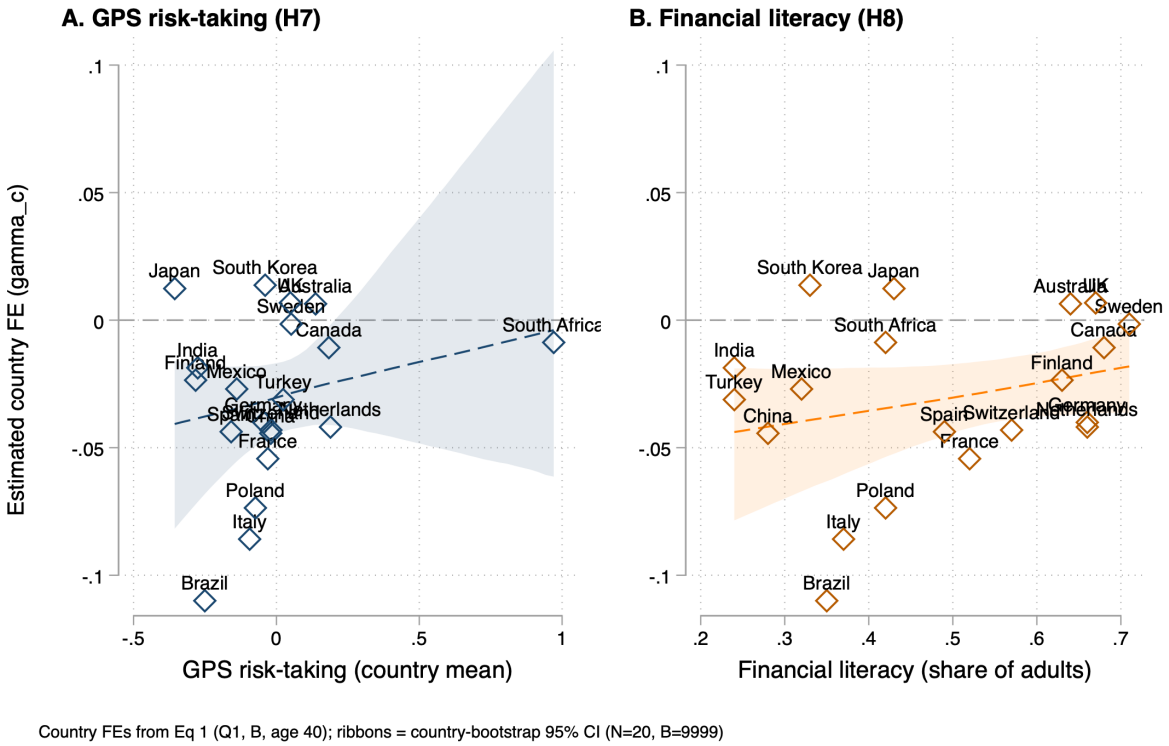
5.8 Does the advice track measured preferences or literacy?

This section asks whether the country effects correlate with two country-level measures a culturally calibrated advisor might respond to. This is a descriptive question rather than a normative one, since the prompt describes a single investor rather than a country mean, and imputing a country's average risk-taking to that individual is an ecological inference the advisor need not make. I use two pre-registered second-stage regressions that estimate whether the country fixed effects $\hat{b}_c$ track the Global Preferences Survey measure of willingness to take risks (Falk et al., 2018) and the share of adults classified as financially literate in the S&P Global FinLit Survey (Klapper et al., 2015). These are distinct from the benchmark calibration slope κ of Section 5.6, which regresses the same country effects on the fundamentals-prescribed deviations rather than on measured preferences or literacy. Literacy predicts equity-market participation at the household level (van Rooij et al., 2011), so advice tailored to the financial sophistication of a country's clientele would recommend more equity where literacy is higher.[7]

Figure 7 plots both relationships. Panel A shows that the advice does not track measured risk preferences. The slope is 0.028 (standard error 0.021), and the wild bootstrap over countries does not reject zero ($p = 0.272$, 95 percent confidence interval $[-0.058, 0.203]$). The point estimate is not only imprecise but economically small. A country one full standard deviation more risk-tolerant than another earns

[7]Each test is a bivariate slope over the twenty non-U.S. countries, with the same inference as the calibration slope: a wild bootstrap over countries (Webb weights, 9,999 replications) as the primary procedure, standard errors alongside, and a two-stage bootstrap that propagates the first-stage uncertainty in $\hat{b}_c$. Like the calibration slope κ, these hypotheses sit at the second altitude of the registration: pre-specified in full, registered as descriptive, with the covariate vectors frozen by a dated amendment before any second-stage coefficient was estimated. Because that amendment post-dates the pilot, I keep these descriptive rather than confirmatory.

Figure 7: Country effects against measured risk preferences and financial literacy



Notes: Estimated country fixed effects $\hat{b}_c$ from equation (1) (Question 1, Version B, age 40, U.S. baseline omitted) plotted against the country mean of the GPS willingness to take risks (Falk et al., 2018) in Panel A and the share of financially literate adults from the S&P Global FinLit Survey (Klapper et al., 2015) in Panel B, for the twenty non-U.S. countries. Ribbons are 95 percent country-bootstrap confidence bands around the fitted bivariate line (9,999 replications).

under three percentage points more equity, and even the widest preference contrast the panel affords, from Japan at −0.36 to South Africa at 0.97, predicts a gap of under four points. The main results on financial advice from LLMs show that Japan receives advice statistically indistinguishable from the U.S. baseline, while South Africa, by far the most risk-tolerant, receives slightly *less* equity than the baseline rather than more.

Panel B tells the same story for literacy. The H8 slope is 0.055 (standard error 0.040, wild bootstrap $p = 0.178$, confidence interval $[-0.030, 0.151]$). Spanning the full literacy range in the panel, from 24 percent of adults in India and Turkey to 71 percent in Sweden, moves predicted advice by under three percentage points. The raw positive association also owes much to composition. The top of the literacy distribution is occupied by the English-prompting countries and the Nordics, whose advice sits at the baseline level, while the bottom mixes countries with large negative effects (Brazil, Italy) and countries with small or no effects (Turkey, South Korea). Once the multivariate regression below accounts for the English-prompt countries, the literacy coefficient changes sign.[8]

[8]The pre-registered diagnostics do not change this reading. Weighting countries by the precision of their first-stage estimates moves the slopes to 0.022 and 0.052. Neither estimate is the artifact of a single country: leave-one-out slopes range from 0.015 (dropping Brazil) to 0.043 (dropping Japan) for risk-taking and from 0.034 to 0.073 for literacy, always positive and always small. The two-stage bootstrap, which resamples the thirteen models and re-estimates the country effects on every draw, produces intervals of $[0.016, 0.040]$ for H7 and $[0.013, 0.099]$ for H8, both excluding zero.

Together with the calibration slope $\hat{\kappa} = 0.13$ of Section 5.6, the second stage completes a consistent picture.[9] The advice does not move with what local fundamentals prescribe, and it does not move with the measured risk preferences or financial literacy of the populations being advised. The cross-country structure that does exist lines up instead with proximity to the language and institutions of the models' training corpus.

5.9 What the simple model leaves out

I note four omissions from the simple model and the likely direction of each. First, the model is static, so with mean-reverting returns longer horizons would justify higher equity shares for the younger investors (Campbell and Viceira, 2002). Second, labor income risk is reduced to a single loading ζ, which lowers $\alpha^I c$ but, in the baseline calibration, does not move it off the corner for a stake this small relative to total wealth. Only the high-ζ cases of Section 5.3 do. Third, taxes on investment returns and housing wealth are excluded from the portfolio benchmarks. Fourth, the benchmark conditions on an average earner, so any country where \$50,000-equivalent savings implies a very different position in the wage distribution (Brazil again) inherits calibration error in Hc and P_c. Each of these would let the advice match a richer benchmark than α_c^S, but the stated-investor arm of Section 6 shows the models do not act on balance-sheet information even when it is stated in the prompt.

None of these caveats enters the computation of Benchmark S, which requires only return moments and preferences. Reassuringly, the one existing full life-cycle treatment of LLM advice reaches the same corner as Benchmark I: Choukhmane et al. (2026) find that a calibrated Cocco et al. (2005) model implies near-universal equity holding for young and middle-aged average earners, and that it cannot rationalize the more conservative shares LLMs actually recommend early in life.

6 The Advice Rule

The country-level results ask whether the same investor receives different advice in different places. The stated-investor arm of Section 3.4 asks the complementary question: which rule generates the advice? The arm traces the benchmark's comparative statics one primitive at a time. Stated risk tolerance targets the risk-aversion coefficient in the denominator of the Merton rule (6). The stated pension targets the outside-wealth multiplier of Benchmark I in equation (9). The stated horizon is, strictly speaking, not in the benchmark, but allows me to test whether models follow the classic result that the optimal share is horizon-irrelevant under CRRA preferences with i.i.d. returns. Because each treatment

[9] The multivariate companion regression, which the registration keeps descriptive with no inferential weight at $N = 20$, regresses the country effects jointly on GPS risk-taking, GPS patience, literacy, and an indicator for the English-prompting countries. Risk-taking, literacy, and patience contribute nothing once the English-prompt countries are accounted for (-0.012, -0.091, and $+0.066$, none distinguishable from zero on HC2 standard errors), while the English-prompt indicator is associated with advice 4.2 percentage points higher (standard error 1.2), and the four covariates together explain 39 percent of the variation in the twenty country effects.

moves one primitive with a sharp theoretical prediction, the sign pattern of the responses discriminates among three candidate models of the advisor: the frictionless Merton planner, the lifecycle planner with background wealth, and the risk questionnaire of popular financial advice.

On the four crossed risk–pension cells the estimating equation is

$$y_{icm} \;=\; \beta_R \,\mathrm{Tol}_i \;+\; \beta_P \,\mathrm{Gen}_i \;+\; \beta_{RP} \,\mathrm{Tol}_i \,\mathrm{Gen}_i \;+\; \phi_c \;+\; \lambda_m \;+\; \varepsilon_{icm}, \tag{14}$$

where Tol_i and Gen_i indicate the risk-tolerant and generous-pension statements, the omitted cell is the risk-averse investor with the limited pension, and ϕ_c and λ_m are country and model fixed effects. On the two horizon cells the analogous equation regresses the recommended share on an indicator for the twenty-year horizon with the one-year horizon omitted, yielding β_H. Inference is the same model-level wild cluster bootstrap as the primary regressions.

There are three main findings from this exercise, which are reported in Appendix Table A5 (the full distribution in each treatment cell) and Appendix Table A6 (the four coefficient estimates). First, the advice responds strongly to stated risk tolerance: the equity share increases by about 46 percentage points for the risk-tolerant investor compared to the risk-averse investor, moving from 0.30 in the risk-averse cells to 0.78 in the risk-tolerant ones. The response is almost identical in all countries, with the response clustered in a band of just 3 percentage points across the 21 countries.

Second, the advice responds strongly to the stated horizon: the one year horizon yields a mean equity share of 0.36, compared to an equity share of 0.76 for the twenty year horizon. The horizon moves the equity share substantially in all countries, but the cross-country span is somewhat larger at 11.9 percentage points, compared to the tight band for the stated risk tolerance.

Third, the advice responds at most weakly to the stated pension. The generous pension lowers the recommended share by 1.0 percentage point at the risk-averse baseline ($\hat{\beta}_P = -1.0$, confidence interval $[-2.3, +0.2]$), and the risk-by-pension interaction of $+3.6$ points ($[2.0, 5.2]$) raises it by about 2.6 points for the risk-tolerant investor. The average marginal effect is about $+0.8$ points, an order of magnitude below the background-wealth adjustment a generous pension should trigger.

These comparative statics identify what kind of advisor an LLM is. The levels of the advice are consistent with a Merton (1969)–Samuelson (1969) planner holding plausible preferences: inverting the Merton rule (6) at the paper's calibration puts the implied relative risk aversion at roughly 2 for the stated-tolerant cells, 5 for the stated-averse cells, and 3 for the no-statement reference, all inside the conventional range. The comparative statics, however, are not those of any frictionless CRRA planner with i.i.d. returns, which predicts exact horizon invariance where the data show a forty-point response. Nor are they those of a lifecycle planner. A stated generous pension is a bond-like annuity claim that should

crowd financial-portfolio equity *in* under any total-wealth model (Cocco et al., 2005), and the estimated response is at most a few points, an order of magnitude too small for such an adjustment. The corner solution of Benchmark I sharpens this reading: an advisor that fully integrated the balance sheet would sit at 100 percent equity in every cell, which the interior levels already rule out, and conditional on interior advice any degree of balance-sheet integration implies $\beta_P > 0$. Mean reversion in returns (Campbell and Viceira, 2002) could rationalize the horizon response but not the pension null. A planner sophisticated enough to price return dynamics into a forty-point horizon effect should not ignore pension wealth entirely.

Instead, the models respond strongly to the two cues that populate retail risk questionnaires—stated risk tolerance and investment horizon, and not at all to the balance-sheet reasoning that normative theory requires. The age analysis points to the same source: the glide path of Section 4.5 is a compressed, language-invariant rendering of the 100-minus-age heuristic, anchored on the industry's 60/40 portfolio at mid-career (Figure 5).[10] Because pension generosity is stated directly in the prompt, no knowledge of local institutions is required to act on it, so the pension null also sharpens the interpretation of the compression results of Section 5.5: a calibration slope near zero could in principle reflect ignorance of local pension systems, but the stated-investor arm shows that the models fail to use pension information even when it is handed to them.

7 Mechanisms in the Explanation Text

Finally, I examine the explanation text that the models supplied with each prompt. The explanations separate two accounts of the compression. Under the first, the model does not know how the investor's country differs from the United States. Under the second, it knows but does not let that knowledge reach the recommended share.

In each prompt the model justifies its recommendation in at most two hundred words, and each free-text rationale is machine-coded into four binary indicators—inflation or real-value protection, an existing pension or retirement entitlement, investment horizon, and local or home-country financial context—by a single out-of-panel judge model (Claude Opus 4.7), shown only the explanation text and blind to the country, model, and allocation that produced it.[11] Three of the four categories pass both gates and enter

[10] The rejection of horizon invariance is not uniform across the panel: two of thirteen models are nearly horizon-invariant, while the most responsive model moves 79 percentage points between the one-year and twenty-year cells. The one-year cell is also where the panel disagrees most, against near-consensus at twenty years. The risk-tolerance response, by contrast, is positive for all thirteen models and remarkably uniform across countries.

[11] The coding instrument was frozen at pre-registration and validated by a panel of four judge models from independent vendor families, none among the thirteen models under study and each shown only the explanation text. On a stratified validation subsample the panel was held to two pre-registered inter-rater reliability gates, reported before any coefficient in this section was estimated and both requiring $\kappa \geq 0.6$: a reproducibility gate on the minimum pairwise Cohen's κ across the four judges and a face-validity gate on the κ between the primary judge and my hand-codes on a bilingual English-and-Swedish anchor. Inflation, pension, and local context clear both gates comfortably (minimum pairwise judge κ of 0.95, 0.95, and 0.87). The primary judge, Claude Opus 4.7, then coded every rationale in the analysis sample, and a deterministic multilingual keyword dictionary, coded independently, agrees with it on 99, 96, and 73 percent of rows,

pre-specified testing.[12]

For each admitted category the country-fixed-effects specification of equation (1), I re-estimated the equation with the binary rationale code in place of the equity share as the outcome (Question 1, Version B, age 40), so that the reasoning and the number are measured by one identical design. The resulting country effects test whether a rationale is invoked at different rates across countries. Inference is the same model-level wild cluster bootstrap as the primary regressions, Holm-corrected within each category's family of twenty contrasts.

The main finding is that the reasoning shown to the user localizes, even if the previous results show that the equity share varies little by country. Local financial context appears in 18.0 percent of the U.S. explanations but in every other country far more often, by 30 to 78 percentage points, with nineteen of the twenty contrasts surviving the Holm correction (Table 3). The rate reaches near-saturation in South Africa, Mexico, and Brazil, where more than nine explanations in ten name a local instrument, tax, or pension system, and it stays close to half even in the countries where it is lowest, South Korea and Sweden at roughly 47 percent. Set against the near-zero response of the allocation to the same localization cues in Section 4, this is the central fact of the section. When the models answer the allocation question outside the United States, they reach for local knowledge in the prose and leave it out of the number.

Inflation reasoning is elevated exactly in the high-inflation economies, rising 56 percentage points in Turkey and 33 in South Africa, both surviving correction, with Poland, Brazil, Mexico, and India also positive though short of the corrected threshold, against a U.S. base of 34 percent. The pension rationale, by contrast, is nearly universal and flat. Almost every explanation names an existing pension or retirement entitlement (97.5 percent in the United States), and no country departs from that rate after correction. The models therefore mention the pension everywhere while, as the stated-investor arm shows, declining to size the allocation to it, so the rationale enters the words but not the number. Horizon reasoning is also present in every explanation and hence untestable for country variation.

Re-estimating the Version-C regression of equation (2) with the local-context code as the outcome, naming local instruments raises the rate of local references by 23.3 percentage points relative to Version B ($\hat{\phi}^{\text{text}} = 0.233$, wild bootstrap $p < 0.001$, confidence interval $[0.143, 0.329]$), a large and precisely estimated response of the reasoning to a cue that moved the allocation itself by -0.7 points (H5, Section 4). Again, the treatment that hands the model the local institutional context outright reshapes what it says and leaves what it recommends essentially unchanged.

Taken together, the user-facing rationales localize while the allocation barely moves. Outside the United States the models name local instruments, taxes, and pension systems far more often, yet the recom-

local context the hardest category by every measure. See Appendix A and the judge-validation amendment of record.

[12] The fourth, horizon, is mentioned in every explanation and thus leaves no variation to test.

Table 3: Rationale composition by country in the explanation text

	Inflation		Pension		Local context	
Country	Est.	Holm p	Est.	Holm p	Est.	Holm p
Turkey	56.3	0.019	0.6	1.000	70.6	<0.001
South Africa	33.2	<0.001	0.0	1.000	78.2	<0.001
Poland	31.4	0.090	−8.2	0.470	56.3	0.003
Brazil	25.1	0.441	−0.3	1.000	75.8	<0.001
Mexico	21.8	0.359	−3.4	0.930	77.2	<0.001
India	17.8	0.803	−0.2	1.000	70.3	<0.001
United Kingdom	5.8	0.141	−8.5	0.128	71.8	<0.001
South Korea	0.2	1.000	−5.5	1.000	29.2	0.059
Canada	−0.2	1.000	2.3	1.000	59.4	<0.001
Australia	−2.8	1.000	−2.6	1.000	68.2	<0.001
Netherlands	−4.7	1.000	−0.5	1.000	43.0	0.009
Italy	−8.8	1.000	−1.4	1.000	70.5	<0.001
Japan	−9.4	1.000	−1.1	1.000	36.3	0.025
Spain	−9.5	1.000	1.9	1.000	61.1	<0.001
Germany	−12.5	1.000	1.5	1.000	45.1	<0.001
Finland	−13.3	0.965	−8.0	0.169	39.4	0.005
France	−14.8	1.000	−0.9	1.000	59.0	<0.001
Switzerland	−16.6	0.998	−3.2	1.000	66.5	<0.001
Sweden	−19.4	0.392	−6.8	1.000	29.6	0.013
China	−20.5	0.333	−2.6	1.000	32.9	0.025

Notes: Country fixed effects from the linear probability model of equation (1), estimated on each binary rationale code in place of the equity share (Question 1, Version B, age 40, $N = 13{,}639$), in percentage points relative to the omitted U.S. baseline. A positive entry means the rationale appears more often than in U.S. explanations. The U.S. baseline rates are 33.7 percent (inflation), 97.5 percent (pension), and 18.0 percent (local context). Holm p-values correct the model-level wild cluster bootstrap p-values (Webb six-point weights, 9,999 replications, $G = 13$) within each category's family of twenty contrasts. Countries are ordered by the inflation estimate. The horizon category is omitted because every explanation invokes it (prevalence 100 percent), leaving its country effects undefined.

mended share is nearly unchanged. The coding verifies that local context is *mentioned*, not that it is accurate or used to set the allocation, so this describes the rationales shown to users rather than the computation behind the number.

The design observes advice rather than model internals, so this remains a reading of explanations given to users rather than a measurement of the computation. However, the picture is consistent with the multilingual-interpretability literature, which shows that a task solved in a shared, English-leaning concept space where translation and local framing rewrite the words of the question and the answer without reaching the policy that produces the number (Wendler et al., 2024; Wu et al., 2025; Schut, Gal and Farquhar, 2025).

8 Conclusion

This paper asks whether LLMs tailor portfolio advice to the country of the investor who asks. A pre-registered experiment poses the same allocation problem to thirteen frontier models across twenty-one countries, decomposes any country effect into language, country-naming, and local-framing components,

and confronts the advice with a country-specific normative benchmark. The headline result is that advice occupies a twelve-point band near whose top the U.S. baseline sits, every country contrast that survives correction is negative and modest, and the language and local-instrument treatments each move the recommended equity share by roughly a point or less. Which model a household consults matters an order of magnitude more than which country it consults from. Moreover, the compression is not an inability to distinguish countries, as the models' explanations reflect local features. However, this reasoning does not lead the model to change their recommendation.

The benchmark and the stated-investor arm together identify what kind of advisor an LLM is. The advice does not track what local fundamentals prescribe, and it does not track the measured risk preferences or financial literacy of the populations being advised. What it does respond to are the two cues a retail risk questionnaire collects. Stated risk tolerance raises the recommended equity share by 46 percentage points, and a stated horizon—irrelevant under CRRA preferences with i.i.d. returns—moves it by 40, while a stated generous pension barely registers. It is the joint pattern, a large horizon response beside a near-absent pension response, that identifies the rule as the "popular" financial advice of Canner et al. (1997) and Choi (2022) rather than a portfolio rule over the household's balance sheet. The near-invariance to language is consistent with interpretability evidence that multilingual models solve the task in a shared, English-leaning concept space, so translation changes the words of the question and the answer but not the policy that produces the advice.

These results bear on a question this paper does not answer by assumption: is uniform advice a defect, or is it what a good advisor should do for investors who differ mainly in preferences a stranger has no business inferring from their nationality? There is a real case for the second reading. If every investor holds the same world-market portfolio, sizing the risky share to preferences rather than to a country name is arguably the more defensible choice. The evidence here does not overturn that reading so much as sit alongside it: the institutional channels the world-index logic says should still move advice, safe rates, currency risk, pension generosity, prescribe a thirty-two-point spread against the twelve the advice spans, and letting preferences vary with measured national risk-taking does not close that gap. The compression runs wider than even a culturally agnostic, world-index advisor would produce. The shared, English-anchored computation of Section 7 is one candidate account of that residual gap, though the design observes advice rather than model internals and cannot cleanly separate it from an advisor that is simply insensitive to institutional detail as well as to preferences.

These findings speak to the market structure that LLM advice is creating. Financial advice was a nationally segmented service, walled in by regulation, by the trust a credence good requires, and by language. The same handful of foundation models now answer the question in every language at negligible marginal cost, outside the regulatory perimeter that bound human advisors. This paper shows that they

answer it with nearly one policy everywhere.

The cost of this compression is small on average and concentrated in its incidence. At plausible calibrations the country gap costs the median household single-digit basis points a year, about forty dollars over five years on the prompt's stake, so a one-size-fits-all advisor is a benign default for most of the panel. The burden lands elsewhere. A single policy anchored near the U.S. training baseline leaves an uncorrected deviation that grows with the distance between local fundamentals and that baseline, and because the welfare loss is quadratic in that deviation, the households whose circumstances least resemble the environment in which the rule was formed bear the largest cost, reaching seven percent of the stake in Turkey while western Europe pays a few basis points at most. Concentration compounds this incidence, because advice drawn from a few foundation models trained on the same corpus produces a common error across the households exposed to it, a correlation the fragmented industry of human advisors would have diluted. The stated-investor arm names the failure behind that error. The advice neglects the investor's balance sheet even when the pension is stated outright, the very margin that suitability and best-interest rules oblige a human advisor to weigh, so the compression sits at once beyond those rules and at odds with their purpose. Whether the regulatory perimeter should extend to advice of this kind is a question this market structure now poses.

Three limitations bound these conclusions and mark the natural extensions. First, the design observes advice rather than model internals, so the reading that translation leaves the underlying policy intact is an inference from behavior consistent with the interpretability literature rather than a direct measurement. Second, the normative benchmark is deliberately static and closed form. A country-calibrated life-cycle model in the tradition of Cocco et al. (2005) would add realism, though the one existing such treatment reaches the same corner as the integrated benchmark here (Choukhmane et al., 2026), so the balance-sheet logic is unlikely to be overturned. Third, the estimates are a snapshot of a fast-moving panel of models collected within a two-day window. Whether the compression documented here narrows or widens as models improve, and whether localization can be induced by retrieval or fine-tuning without reintroducing the language-keyed inconsistency the shared concept space currently avoids, are the questions a borderless and concentrated advice market makes first-order.

References

Bagliano, Fabio C., Carolina Fugazza, and Giovanna Nicodano, "Life-cycle welfare losses from rules-of-thumb asset allocation," *Economics Letters*, 2021, *198*, 109655. 29

Bodie, Zvi, Robert C Merton, and William F Samuelson, "Labor supply flexibility and portfolio choice in a life cycle model," *Journal of economic dynamics and control*, 1992, *16* (3-4), 427–449. 25

Bolukbasi, Tolga, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam T. Kalai, "Man is to computer programmer as woman is to homemaker? Debiasing word embeddings," in "Advances in Neural Information Processing Systems," Vol. 29 2016. 6

Campbell, John Y and Luis M Viceira, *Strategic asset allocation: portfolio choice for long-term investors*, Clarendon Lectures in Economic, 2002. 25, 34, 36

Canner, Niko, N. Gregory Mankiw, and David N. Weil, "An asset allocation puzzle," *American Economic Review*, 1997, *87* (1), 181–191. 4, 39

Choi, James J, "Popular personal financial advice versus the professors," *Journal of Economic Perspectives*, 2022, *36* (4), 167–192. 4, 39

Choukhmane, Taha, Tim de Silva, Weidong Lin, and Matthew Akuzawa, "AI Financial Advice: Supply, Demand, and Life Cycle Implications," *Available at SSRN*, 2026. 3, 5, 21, 23, 24, 28, 34, 40

Christelis, Dimitris, Dimitris Georgarakos, Tullio Jappelli, and Maarten van Rooij, "Consumption uncertainty and precautionary saving," *Available at SSRN 2698098*, 2015. 6

Cocco, João F., Francisco J. Gomes, and Pascal J. Maenhout, "Consumption and portfolio choice over the life cycle," *Review of Financial Studies*, 2005, *18* (2), 491–533. 23, 34, 36, 40

D'Acunto, Francesco and Alberto G. Rossi, "Robo-advice: Transforming households into rational economic agents," *Annual Review of Financial Economics*, 2023, *15*, 543–563. 6

_ , Nagpurnanand Prabhala, and Alberto G. Rossi, "The promises and pitfalls of robo-advising," *Review of Financial Studies*, 2019, *32* (5), 1983–2020. 5

Dimson, Elroy, Paul Marsh, and Mike Staunton, *Triumph of the Optimists: 101 Years of Global Investment Returns*, Princeton, NJ: Princeton University Press, 2002. 57

Falk, Armin, Anke Becker, Thomas Dohmen, Benjamin Enke, David Huffman, and Uwe Sunde, "Global evidence on economic preferences," *Quarterly Journal of Economics*, 2018, *133* (4), 1645–1692. 3, 6, 7, 27, 32, 33, 57

Financial Health Network, "Financial Health Pulse 2025 U.S. Trends Report," Technical Report, Financial Health Network 2025. 1

Foerster, Stephen, Juhani T Linnainmaa, Brian T Melzer, and Alessandro Previtero, "Retail financial advice: does one size fit all?," *The Journal of Finance*, 2017, *72* (4), 1441–1482. 4

Foltyn, Richard and Jonna Olsson, "The Worth of a "Wo": Gender Bias in Financial Advice from LLMs," Working Paper, Norwegian School of Economics 2026. 4, 6, 15, 21

Fuster, Andreas, Paul Goldsmith-Pinkham, Tarun Ramadorai, and Ansgar Walther, "Predictably unequal? The effects of machine learning on credit markets," *Journal of Finance*, 2022, *77* (1), 5–47. 6

Gálvez, Julio and Gonzalo Paz-Pardo, "Richer earnings dynamics, consumption and portfolio choice over the life cycle," *Journal of Financial Economics*, 2026, *176*, 104206. 25, 26

Guiso, Luigi and Paolo Sodini, "Household finance: An emerging field," in George M. Constantinides, Milton Harris, and René M. Stulz, eds., *Handbook of the Economics of Finance*, Vol. 2, Elsevier, 2013, pp. 1397–1532. 6

J.D. Power, "As More U.S. Consumers Struggle with Rising Prices, Many Turn to Artificial Intelligence for Financial Advice," Banking and Payments Intelligence Report — Press Release, J.D. Power August 2025. Based on a survey of approximately 4,000 U.S. consumers. 1

Jordà, Òscar, Katharina Knoll, Dmitry Kuvshinov, Moritz Schularick, and Alan M. Taylor, "The rate of return on everything, 1870–2015," *Quarterly Journal of Economics*, 2019, *134* (3), 1225–1298. 57

Kassner, Nora, Philipp Dufter, and Hinrich Schütze, "Multilingual LAMA: Investigating Knowledge in Multilingual Pretrained Language Models," in "Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics" Association for Computational Linguistics 2021. 18

Khemka, Gaurav, Mogens Steffensen, and Geoffrey J. Warren, "How sub-optimal are age-based life-cycle investment products?," *International Review of Financial Analysis*, 2021, *73*, 101619. 29

Klapper, Leora, Annamaria Lusardi, and Peter van Oudheusden, "Financial Literacy Around the World: Insights from the Standard & Poor's Ratings Services Global Financial Literacy Survey," Working Paper, Global Financial Literacy Excellence Center 2015. 3, 32, 33

Li, Ang Leo, Fujing Xue, Zixuan Zeng, and Xiaofeng Zhao, "Stereotypes in the Machine: Gender and Racial Bias in LLM-Generated Financial Advice," *Available at SSRN 7069438*, 2026. 4

Lindsey, Jack, Wes Gurnee, Emmanuel Ameisen, Brian Chen, Adam Pearce, Nicholas L. Turner, Craig Citro et al., "On the Biology of a Large Language Model," Transformer Circuits Thread, Anthropic 2025. 2, 17

Lloyds Banking Group, "Over 28 Million Adults Now Using AI Tools to Help Manage Their Money," Press Release, Lloyds Banking Group November 2025. 1

Love, David A., "Optimal rules of thumb for consumption and portfolio choice," *The Economic Journal*, 2013, *123* (571), 932–961. 21, 29

Merton, Robert C., "Lifetime portfolio selection under uncertainty: The continuous-time case," *Review of Economics and Statistics*, 1969, *51* (3), 247–257. 3, 23, 35

Navigli, Roberto, Simone Conia, and Björn Ross, "Biases in large language models: Origins, inventory, and discussion," *ACM Journal of Data and Information Quality*, 2023, *15* (2). 6

OECD, *Pensions at a Glance 2023: OECD and G20 Indicators*, Paris: OECD Publishing, 2023. 57

Qi, Jirui, Raquel Fernández, and Arianna Bisazza, "Cross-lingual consistency of factual knowledge in multilingual language models," in "Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing" 2023, pp. 10650–10666. 18

Reuter, Jonathan and Antoinette Schoar, "Demand-side and supply-side constraints in the market for financial advice," *Annual Review of Financial Economics*, 2024, *16*, 229–255. 6

Ross, Jillian and Andrew W Lo, "One Size Fits None: Heuristic Collapse in LLM Investment Advice," *arXiv preprint arXiv:2604.23837*, 2026. 4, 5

Rossi, Alberto G. and Stephen P. Utkus, "The diversification and welfare effects of robo-advising," *Journal of Financial Economics*, 2024, *157*, 103869. 5

Samuelson, Paul A., "Lifetime portfolio selection by dynamic stochastic programming," *Review of Economics and Statistics*, 1969, *51* (3), 239–246. 23, 35

Schut, Lisa, Yarin Gal, and Sebastian Farquhar, "Do Multilingual LLMs Think in English?," *arXiv preprint arXiv:2502.15603*, 2025. 38

Tang, Tianyi, Wenyang Luo, Haoyang Huang, Dongdong Zhang, Xiaolei Wang, Wayne Xin Zhao, Furu Wei, and Ji-Rong Wen, "Language-Specific Neurons: The Key to Multilingual Capabilities in Large Language Models," in "Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)" Association for Computational Linguistics Bangkok, Thailand 2024, pp. 5701–5715. 2, 17

van Rooij, Maarten, Annamaria Lusardi, and Rob Alessie, "Financial literacy and stock market participation," *Journal of Financial Economics*, 2011, *101* (2), 449–472. 6, 32

Wendler, Chris, Veniamin Veselovsky, Giovanni Monea, and Robert West, "Do llamas work in english? on the latent language of multilingual transformers," in "Proceedings of the 62nd Annual

Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)" 2024, pp. 15366–15394. 2, 17, 38

Wu, Zhaofeng, Xinyan Yu, Dani Yogatama, Jiasen Lu, and Yoon Kim, "The semantic hub hypothesis: Language models share semantic representations across languages and modalities," in "International Conference on Learning Representations," Vol. 2025 2025, pp. 53705–53723. 2, 17, 38

A Appendix Tables and Figures

Table A1: The country panel: prompt language, investment amount, and local financial context

			Local financial context (Version C)		
Country	Language	Amount	Equity instrument	Government bonds	Pension reference
United States	English	USD 50,000	Broad US total-market index fund (Vanguard or Fidelity)	US Treasury bonds	401(k) through employer
United Kingdom	English	GBP 40,000	FTSE All-World ETF	UK Gilts	Workplace pension through employer
Australia	English	AUD 75,000	ASX-listed global index ETF	Australian government bonds	Superannuation fund through employer
Canada	English	CAD 70,000	Global index ETF	Government of Canada bonds	Workplace pension plus RRSP
India	English	INR 4,000,000	Nifty index fund	Government of India bonds	EPF plus NPS account
South Africa	English	ZAR 900,000	Global index ETF (MSCI World)	South African government bonds	Pension fund through employer
Sweden	Swedish	SEK 500,000	Global index fund (via Avanza or Nordnet)	Swedish government bonds	*Tjänstepension* through employer
Finland	Finnish	EUR 45,000	World index fund (MSCI World, via Nordnet or OP)	Finnish government bonds	*Työeläke* (TyEL) through employer
Germany	German	EUR 45,000	MSCI World ETF	*Bundesanleihen*	*Betriebliche Altersvorsorge* (bAV)
France	French	EUR 45,000	World equity ETF (MSCI World)	OATs	*Plan d'épargne retraite* (PER)
Spain	Spanish	EUR 45,000	World index fund (MSCI World)	Spanish government bonds	*Plan de pensiones de empresa*
Japan	Japanese	JPY 7,500,000	Global equity index fund (eMAXIS Slim All Country)	JGBs	iDeCo plus national pension (*kokumin nenkin*)
Brazil	Portuguese (BR)	BRL 250,000	Global-ETF BDRs (BURT39, BWOR39)	Tesouro Direto	Private pension plan (PGBL/VGBL)
Netherlands	Dutch	EUR 45,000	World index ETF (MSCI World)	Dutch government bonds	*Pensioenfonds* through employer
Italy	Italian	EUR 45,000	World index ETF (MSCI World)	BTPs	*Fondo pensione* through employer
Switzerland	German (CH)	CHF 45,000	MSCI World ETF	Swiss government bonds	*Pensionskasse* (second pillar) plus *Säule 3a*
Poland	Polish	PLN 200,000	World index fund (MSCI World ETF)	Polish government bonds	PPK plan through employer
South Korea	Korean	KRW 70,000,000	Global equity index fund (KODEX or TIGER world ETF)	KTBs	Retirement pension (*toejik yeongeum*) plus national pension (*gukmin yeongeum*)
China	Chinese (simplified)	CNY 350,000	QDII global index fund (MSCI World)	Chinese government bonds	Basic pension insurance (*jiben yanglao baoxian*) plus enterprise annuity
Mexico	Spanish (MX)	MXN 1,000,000	Global index fund (MSCI World ETF)	CETES	Afore retirement account
Turkey	Turkish	TRY 2,000,000	Global equity index fund (MSCI World ETF)	Turkish government bonds	BES private pension account

Notes: The upper block lists the six English-prompting countries, which have no Version B-cross cell, and the lower block the fifteen local-language countries. The investment amount is the round local-currency equivalent of $50,000, identical across Versions B, C, and B-cross within a country, and rendered in the prompt in the native number format (e.g. "500 000 kr" in Sweden, "45.000 Euro" in Germany). The three context columns give the country-specific vocabulary that Version C substitutes for the generic Version B wording ("a broad stock market index fund," "government bonds," "a standard pension plan"), condensed from the exact prompt sentences. In most countries the Version C equity instrument is a global or MSCI World-type tracker. India is the exception, naming a domestic Nifty index fund, so for India the Version C step also shifts the underlying equity exposure and not only its label. Pension terms that appear in Japanese, Korean, or Chinese script in the prompts are shown here in romanized form. The full prompt texts for every country, version, and question are provided in the replication package.

Table A2: Query accounting by experimental arm

Arm	Version	Countries	Question–age blocks	Cells	Queries
Country/language	B	21	Q1 at ages 30/40/55	63	40,950
	C	21	Q1 at ages 30/40/55	63	40,950
	B-cross	15	Q1 at ages 30/40/55	45	29,250
	B	21	Q2 and Q3 at age 40	42	27,300
	C	21	Q2 and Q3 at age 40	42	27,300
	B-cross	15	Q2 and Q3 at age 40	30	19,500
Stated investor	B$\gamma\rho\tau$ (six cells)	21	Q1 at age 40	126	81,900
Total				411	267,150

Notes: A cell is one country × version × question × age combination, and for the stated-investor arm one country × treatment cell. Every cell is queried by all thirteen models fifty times, so each cell contributes $13 \times 50 = 650$ queries. Version B-cross exists only for the fifteen non-English countries. The two pre-registered collection-robustness cells (temperature bookends at 0 and 1.0, and Version B with the investment amount stated in U.S. dollars) added 8,450 queries outside this grid and are reported in Section 4.6.

Table A3: Advice by model: level, dispersion, and stability

				Stability across repetitions	
Model	Vendor	Mean	SD	Within-cell SD	Identical cells
Gemini 3.1 Pro Preview	Google	0.37	0.08	0.037	0.05
GPT-5-mini	OpenAI	0.38	0.08	0.043	0.00
GPT-5.5	OpenAI	0.41	0.05	0.032	0.00
DeepSeek V4 Flash	DeepSeek	0.45	0.12	0.110	0.00
DeepSeek V4 Pro	DeepSeek	0.48	0.12	0.105	0.00
Grok 4.3	xAI	0.54	0.09	0.065	0.00
Gemini 3 Flash Preview	Google	0.56	0.07	0.000	0.62
Gemini 3.5 Flash	Google	0.57	0.06	0.020	0.24
Claude 4.8 Opus	Anthropic	0.58	0.04	0.000	0.52
Claude 4.6 Sonnet	Anthropic	0.63	0.04	0.000	0.67
Claude 4.6 Opus	Anthropic	0.65	0.07	0.000	0.95
Mistral Medium 3.5	Mistral	0.70	0.04	0.024	0.29
Claude 4.5 Haiku	Anthropic	0.71	0.04	0.007	0.48

Notes: Recommended equity share by model on the primary cell (Question 1, Version B, age 40), pooled across the twenty-one countries, roughly 1,050 responses per model. Within-cell SD is the median, across a model's twenty-one country cells, of the standard deviation over the fifty repetitions in the cell. Identical cells is the share of those cells whose fifty repetitions return one identical number, despite sampling at temperature 0.7. A variance decomposition of the primary cell attributes 69.5 percent of the total variation to model identity, 5.7 percent to the country, 9.5 percent to their interaction, and 15.3 percent to within-cell sampling.

Table A4: Primary-instrument recommendation on the inflation question by country (Question 3, Version B, age 40)

Country	Equities	Infl.-linked	Gold	Real estate	Other
Brazil	13.1	86.7	–	–	0.2
Turkey	17.5	66.0	16.3	–	0.2
Mexico	41.9	58.1	–	–	–
Poland	61.1	38.6	–	–	0.3
South Africa	82.0	16.3	–	–	1.7
India	88.5	7.8	3.7	–	–
Italy	93.2	6.8	–	–	–
South Korea	94.0	6.0	–	–	–
Spain	96.4	3.6	–	–	–
USA	96.8	2.9	–	–	0.3
Canada	97.2	2.6	–	–	0.2
France	97.8	2.2	–	–	–
United Kingdom	98.0	2.0	–	–	–
Australia	98.3	1.5	–	–	0.2
Switzerland	95.2	0.5	–	4.3	–
Netherlands	99.5	0.3	–	–	0.2
Japan	99.7	0.2	–	–	0.2
Sweden	99.8	0.2	–	–	–
China	96.3	0.2	–	–	3.5
Germany	88.6	0.2	–	11.2	–
Finland	100.0	–	–	–	–

Notes: Share of responses (in percent) recommending each instrument category as the primary way to protect savings against inflation (Question 3, Version B, age 40), pooled across the thirteen models. Each response is assigned a single primary instrument by a deterministic multilingual keyword classifier, and the five categories sum to 100; "Other" collects mixed, bond, and money-market funds and the rare response with no reported instrument. Countries are ordered by the combined inflation-protection share (inflation-linked bonds plus gold), descending. A dash denotes a category recommended in under 0.05 percent of responses. This is the categorical companion to the numeric outcome of Table A9, discussed in Section 4.4.

Table A5: Summary statistics on the stated-investor arm by treatment cell

Treatment cell	Mean	SD	p10	Median	p90	N
Risk-averse, limited pension	0.30	0.08	0.20	0.30	0.40	13646
Risk-averse, generous pension	0.29	0.07	0.20	0.30	0.40	13647
Risk-tolerant, limited pension	0.77	0.10	0.60	0.80	0.85	13643
Risk-tolerant, generous pension	0.79	0.09	0.70	0.80	0.90	13648
Horizon 1 year	0.36	0.24	0.00	0.30	0.70	13645
Horizon 20 years	0.76	0.05	0.70	0.75	0.80	13647

Notes: Distribution of the recommended equity share in the stated-investor arm (Version B$\gamma\rho\tau$, Question 1, age 40, English), pooled across the twenty-one countries and thirteen models. The upper block reports the four crossed cells that state the investor's risk attitude and pension generosity at a common five-year horizon, and the lower block the two cells that state only the investment horizon with no risk or pension statement. Each cell is queried by all thirteen models fifty times in each of the twenty-one countries. These summaries describe the sample behind the pre-registered stated-investor tests (H11–H13) in Section 6: the near-vertical response to stated risk attitude, the flat response to stated pension generosity, and the large response to horizon that the CRRA benchmark predicts should be absent.

Table A6: Stated-investor arm: estimated effects of the risk, pension, and horizon statements

	Contrast	Estimate	95% CI	p (wild)
H11	Risk-tolerant vs. risk-averse	46.3	[41.1, 51.4]	<0.001
H12	Generous vs. limited pension	−1.0	[−2.3, 0.2]	0.107
–	Risk × pension interaction	3.6	[2.0, 5.2]	<0.001
H13	20-year vs. 1-year horizon	39.9	[25.7, 53.4]	<0.001

Notes: Coefficients from the pre-registered stated-investor regressions (Version B$\gamma\rho\tau$, Question 1, age 40), estimated with country and model fixed effects on the sample summarized in Table A5. β_R, β_P, and the interaction come from equation (14) on the four crossed risk–pension cells, with the risk-averse, limited-pension cell as the omitted baseline. β_H comes from the horizon analog on the two horizon cells, with the one-year horizon as the baseline. Estimates and confidence intervals are in percentage points of the recommended equity share. p (wild) is the wild cluster bootstrap p-value (Webb six-point weights, 9,999 replications, clustered by model, $G = 13$), and the confidence interval comes from the same procedure. The risk–pension interaction is descriptive and not a pre-registered test. These are the estimates behind hypotheses H11–H13 in Table A13 and Section 6.

Table A7: Country effects on the recommended equity share (Question 1, Version B, age 40)

Country	Estimate	95% CI	p (wild)	p (Holm)
Brazil	−11.0	[−15.0, −6.9]	<0.001	0.002
Italy	−8.6	[−11.9, −5.3]	<0.001	<0.001
Poland	−7.4	[−10.8, −3.9]	<0.001	0.011
France	−5.4	[−7.5, −3.4]	<0.001	<0.001
China	−4.4	[−8.8, −0.2]	0.041	0.371
Spain	−4.4	[−6.3, −2.5]	<0.001	<0.001
Switzerland	−4.3	[−7.6, −1.1]	0.016	0.200
Netherlands	−4.2	[−6.8, −1.7]	0.004	0.049
Germany	−4.0	[−7.7, −0.4]	0.034	0.335
Turkey	−3.1	[−6.5, −0.1]	0.044	0.371
Mexico	−2.7	[−4.2, −1.2]	<0.001	0.003
Finland	−2.4	[−5.0, 0.4]	0.086	0.602
India	−1.9	[−3.3, −0.4]	0.015	0.200
Canada	−1.1	[−2.1, −0.1]	0.029	0.317
South Africa	−0.9	[−2.4, 0.5]	0.219	1.000
Sweden	−0.2	[−3.7, 3.4]	0.931	1.000
Australia	0.6	[−0.7, 1.9]	0.343	1.000
United Kingdom	0.7	[−0.8, 2.2]	0.352	1.000
Japan	1.2	[−2.1, 4.8]	0.492	1.000
South Korea	1.4	[−2.8, 5.7]	0.496	1.000

Notes: Country fixed effects γ_c from equation (1), estimated on the primary cell (Question 1, Version B, age 40, $N = 13{,}639$) with model fixed effects and the United States as the omitted baseline. Estimates and confidence intervals are in percentage points of the recommended equity share, and countries are ordered by point estimate. p (wild) is the wild cluster bootstrap p-value (Webb six-point weights, 9,999 replications, clustered by model, $G = 13$), and the 95 percent confidence interval comes from the same procedure. p (Holm) applies the pre-registered Holm step-down correction across the family of twenty contrasts. Figure 2 plots these estimates.

Table A8: Country effects on the mortgage question (Question 2, Version B, age 40)

Country	Estimate	95% CI	p (wild)	p (Holm)
Japan	−28.6	[−42.5, −15.0]	<0.001	0.002
China	−22.7	[−32.5, −13.7]	<0.001	0.002
Switzerland	−21.6	[−29.8, −14.0]	<0.001	0.002
Germany	−21.1	[−30.0, −12.0]	<0.001	0.002
Spain	−20.4	[−28.6, −12.3]	<0.001	0.002
Italy	−19.5	[−26.7, −12.5]	<0.001	0.002
Finland	−19.1	[−27.1, −11.3]	<0.001	0.002
South Korea	−17.3	[−24.4, −10.1]	<0.001	0.002
Australia	−17.1	[−24.8, −9.8]	<0.001	0.003
Poland	−16.2	[−21.0, −11.5]	<0.001	0.002
Netherlands	−16.0	[−23.2, −9.9]	<0.001	0.002
France	−15.7	[−24.6, −6.6]	<0.001	0.003
United Kingdom	−14.3	[−21.5, −8.1]	<0.001	0.003
Canada	−12.3	[−18.4, −6.6]	<0.001	0.003
Sweden	−10.0	[−16.5, −4.1]	0.001	0.006
Premise-acceptance group				
India	−3.8	[−6.7, −0.8]	0.015	0.060
South Africa	−2.4	[−4.3, −0.7]	0.002	0.008
Mexico	−1.9	[−5.3, 1.7]	0.262	0.787
Brazil	−1.7	[−6.2, 3.0]	0.452	0.904
Turkey	−1.6	[−8.4, 5.3]	0.628	0.904

Notes: Country fixed effects on the recommended fraction to invest in a broad stock market index fund rather than prepay a 4 percent mortgage (Question 2, Version B, age 40, $N = 13{,}619$), estimated with the specification of equation (1) with model fixed effects and the United States as the omitted baseline. Estimates and confidence intervals are in percentage points, and countries are ordered by point estimate within group. Inference is the model-level wild cluster bootstrap (Webb six-point weights, 9,999 replications, $G = 13$), and p (Holm) applies the step-down correction across the twenty contrasts, which form their own family separate from Question 1 as pre-registered. The five countries below the separator are the pre-specified premise-acceptance group, whose prevailing mortgage rate at the filing date exceeded twice the prompted 4 percent, so their contrasts test whether the model accepts or corrects the premise rather than whether it calibrates to local fundamentals. Figure 4 plots these estimates alongside those of the other two questions.

Table A9: Country effects on the inflation question (Question 3, Version B, age 40)

Country	Estimate	95% CI	p (wild)	p (Holm)
Japan	−11.3	[−13.7, −8.9]	<0.001	<0.001
China	−7.2	[−9.6, −4.7]	<0.001	<0.001
South Korea	−6.9	[−10.3, −3.5]	0.001	0.020
Netherlands	−6.5	[−8.9, −4.2]	<0.001	0.002
Switzerland	−5.0	[−7.3, −2.7]	<0.001	0.006
Italy	−5.0	[−7.3, −2.8]	<0.001	0.002
France	−4.3	[−6.9, −1.8]	0.003	0.041
Spain	−4.0	[−6.6, −1.3]	0.007	0.091
Finland	−2.5	[−5.2, −0.1]	0.041	0.473
Germany	−2.0	[−4.1, −0.1]	0.039	0.473
Poland	−1.8	[−5.3, 1.6]	0.269	1.000
United Kingdom	−1.5	[−3.2, 0.1]	0.064	0.636
Turkey	−0.9	[−4.5, 2.7]	0.596	1.000
Australia	−0.2	[−2.2, 1.9]	0.862	1.000
Sweden	−0.1	[−3.1, 2.7]	0.930	1.000
Mexico	0.1	[−3.3, 3.4]	0.973	1.000
India	0.8	[−2.5, 3.9]	0.618	1.000
Canada	1.6	[−0.4, 3.6]	0.100	0.799
South Africa	2.3	[−0.3, 4.9]	0.082	0.739
Brazil	2.6	[−1.0, 5.9]	0.139	0.972

Notes: Country fixed effects on the recommended fraction of savings to move out of a 1 percent savings account in the face of 3 percent inflation (Question 3, the registration's Q4, Version B, age 40, N = 13,637), estimated with the specification of equation (1) with model fixed effects and the United States as the omitted baseline. Estimates and confidence intervals are in percentage points, and countries are ordered by point estimate. Inference is the model-level wild cluster bootstrap (Webb six-point weights, 9,999 replications, G = 13), and p (Holm) applies the step-down correction across the twenty contrasts, which form their own family. The numeric fraction is the primary Question 3 outcome because the collection produced no refusals and the pre-specified refusal contingency never bound (Section 3.7). Figure 4 plots these estimates alongside those of the other two questions, and Table A4 reports the categorical companion.

Table A10: The age gradient and its interaction with pension generosity (Question 1, Version B)

	(1) No interaction	(2) Continuous $\tilde{\rho}$	(3) Median split
Age 30 (vs. 40)	6.24	6.24	6.24
	(0.57)	(0.57)	(0.57)
Age 55 (vs. 40)	-11.13	-11.13	-11.34
	(1.90)	(1.90)	(1.96)
× std. replacement rate (ψ)		-0.02	
		(0.33)	
× above-median pension			0.44
			(0.61)
Country fixed effects	Yes	Yes	Yes
Model fixed effects	Yes	Yes	Yes
Wild p: Age 30	0.000	0.000	0.000
Wild p: Age 55	0.000	0.000	0.000
Wild p: interaction		0.964	0.489
Observations	40,910	40,910	40,910
Model clusters (G)	13	13	13

Notes: Estimates of equation (4) on Question 1, Version B, pooled across the three prompted ages (30, 40, 55), the twenty-one countries, and the thirteen models. The outcome is the recommended equity share in percentage points, and every column includes country and model fixed effects with the United States and age 40 as the omitted baselines. Column (1) fits the age effects alone. Column (2) adds the pre-registered interaction of the age-55 indicator with the standardized OECD net pension replacement rate $\tilde{\rho}_c$ (the coefficient ψ of hypothesis H6). Column (3) replaces it with the coarser interaction of age 55 with an above-median replacement-rate indicator. Parentheses report CR1 cluster-robust standard errors clustered by model. The wild cluster bootstrap p-values (Webb six-point weights, 9,999 replications, clustered by model, $G = 13$) are the inference of record and are reported in the lower panel.

Table A11: Collection-robustness cells: temperature bookend and numeral-magnitude test

Panel A: Country contrasts vs. the U.S. at the temperature bookends

	$T = 0$		$T = 0.7$ (design)		$T = 1.0$	
Country	Estimate	95% CI	Estimate	95% CI	Estimate	95% CI
Brazil	−10.0	[−14.2, −5.8]	−11.0	[−15.0, −6.9]	−11.4	[−15.1, −7.6]
Germany	−3.6	[−7.4, 0.2]	−4.0	[−7.7, −0.4]	−4.2	[−7.7, −0.7]
Sweden	−0.3	[−3.7, 3.1]	−0.2	[−3.7, 3.4]	−0.7	[−4.0, 2.5]
Japan	2.5	[−2.0, 7.0]	1.2	[−2.1, 4.8]	0.2	[−3.4, 4.0]
Mean within-cell SD	0.035		0.038		0.041	

Panel B: Stating the amount as $50,000 instead of the native-currency figure

Country	Mean (Version B)	Mean (B-USD)	Shift	95% CI	p (wild)
Germany	0.53	0.54	1.2	[−0.4, 2.7]	0.130
Japan	0.58	0.55	−2.8	[−5.4, −0.4]	0.010
Sweden	0.57	0.58	1.2	[−0.8, 3.4]	0.255

Notes: Results from the two pre-registered collection-robustness cells (Section 4.6), re-collected on July 6, 2026 alongside the main corpus (Section 3.7). Panel A re-estimates the country fixed effects of equation (1) separately at each sampling temperature on the five bookend countries (Question 1, Version B, age 40, United States omitted), in percentage points of the recommended equity share. The $T = 0.7$ column uses the primary-collection cells and the bookend columns use the re-collected cells, so the $T = 0.7$ estimates differ from Appendix Table A7 only in the estimation sample being restricted to these five countries. The dispersion row reports the mean across country–model cells of the within-cell standard deviation of the recommended share, in fraction units. Panel B compares Version B, whose stake is a round native-currency amount, with an otherwise identical prompt stating the amount as $50,000, for the three pre-registered countries with country and language held fixed. The mean columns are recommended equity shares, and the shift is the B-USD minus Version-B difference in percentage points, estimated with model fixed effects. All p-values and confidence intervals come from the wild cluster bootstrap (Webb six-point weights, 9,999 replications, clustered by model, $G = 13$). The pooled estimates quoted in Section 4.6 average over the countries shown here.

Table A12: Country-specific calibration inputs for the normative benchmark

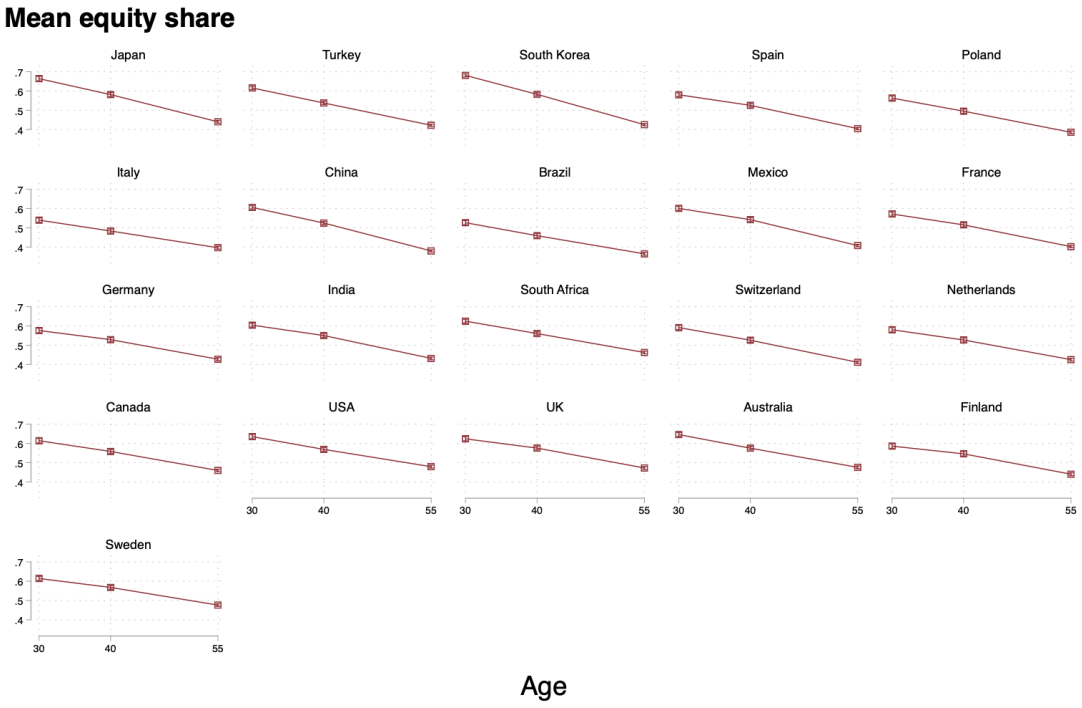
Country	σ_c (%)	r_c (%)	z_c (std)	ρ_c (%)	$\bar{w}_c$ (000s)	R_c (age)	L_c (yrs)	g_c (%)	τ_c^m	π_c (%)	α_c^S ($\bar{\gamma}$=3)
USA (USD)	18.4	0.4	0.12	51.3	87	67	18	1.1	0.15	2.5	0.52
United Kingdom (GBP)	19.1	-0.3	0.05	54.2	47	68	17	0.3	0.00	2.9	0.49
Sweden (SEK)	19.2	-0.3	0.05	66.3	546	70	16	1.0	0.30	2.1	0.48
Finland (EUR)	18.5	-0.1	-0.28	65.7	51	68	17	0.5	0.00	2.0	0.52
Germany (EUR)	18.5	-0.6	-0.04	53.3	55	67	18	0.8	0.00	2.2	0.52
France (EUR)	18.5	0.2	-0.03	70.0	44	65	22	0.6	0.00	1.9	0.52
Spain (EUR)	18.5	0.9	-0.16	86.3	35	65	22	0.5	0.00	2.0	0.52
Japan (JPY)	20.4	-0.1	-0.36	42.4	5,186	65	22	-0.2	0.10	0.9	0.43
Brazil (BRL)	27.5	0.8	-0.25	97.5	56	64	20	1.2	0.00	5.5	0.23
Australia (AUD)	19.0	0.9	0.14	53.0	105	67	20	0.7	0.00	2.7	0.49
Canada (CAD)	18.5	0.3	0.18	45.1	86	65	21	0.9	0.00	2.2	0.52
India (INR)	19.6	1.0	-0.28	44.6	273	58	23	4.9	0.10	6.4	0.46
South Africa (ZAR)	19.0	3.7	0.97	8.9	339	60	20	0.1	0.00	5.5	0.49
Netherlands (EUR)	18.5	-0.5	0.19	96.0	61	70	15	0.1	0.35	2.4	0.52
Italy (EUR)	18.5	1.3	-0.09	79.0	35	70	16	-0.3	0.00	2.0	0.52
Switzerland (CHF)	18.7	0.2	-0.02	47.5	98	65	22	0.7	0.20	0.5	0.51
Poland (PLN)	22.5	0.7	-0.07	40.6	101	62	21	2.6	0.00	3.7	0.35
South Korea (KRW)	18.4	1.0	-0.04	38.9	53,916	65	21	1.3	0.05	2.2	0.53
China (CNY)	19.1	1.4	-0.02	103.6	106	60	22	6.8	0.05	2.2	0.48
Mexico (MXN)	21.7	2.9	-0.14	79.6	277	65	18	0.1	0.00	4.4	0.38
Turkey (TRY)	29.1	-4.2	0.02	96.4	743	64	18	11.5	0.00	19.4	0.21

Notes: Actual country-level inputs to the normative benchmark of Section 5, one row per country. σ_c is the annualized volatility of a world equity index measured in local currency, and r_c is the real safe rate, local government yields less expected inflation, taken as a trailing twenty-year mean. The equity premium $\mu_c - r_c$ is held common across countries at 5.32 percent, the arithmetic mean of the net MSCI World index in U.S. dollars over the U.S. long-term rate (1999–2025), so under common preferences the cross-country variation in α_c^S enters through σ_c alone. z_c is the country-mean GPS willingness to take risks, standardized at the individual level (Falk et al., 2018). ρ_c is the OECD net pension replacement rate for the average earner and $\bar{w}_c$ is the average annual wage in thousands of local currency (OECD, 2023). R_c is the normal retirement age, L_c the years of life expectancy at retirement, and g_c the real wage growth. τ_c^m is the assumed effective mortgage-interest deductibility for the mortgage question, which is income- and itemization-dependent rather than point-sourced, so α_c^{Q2} is reported over a $\tau^m \in [0, 0.5]$ band. π_c is expected inflation from central-bank targets and consensus forecasts. The final column is the derived standalone benchmark share α_c^S at the baseline grid point ($\bar{\gamma} = 3$, $\eta = 0$, world index in local currency), which reproduces the benchmark column of Table 2 and ranges from 0.21 in Turkey to 0.53 in South Korea. Return and wage moments are unhedged local-currency values, with a currency-hedged variant as a robustness row. Standard errors are not separately estimated for each input. The released benchmark output instead spans the full sensitivity grid $\bar{\gamma} \in \{2, 3, 5, 10\}$, $\eta \in \{0, 0.3, 0.6\}$, $\zeta \in \{0, 0.2\}$, and the thin-data group (Brazil, China, India, Turkey, Mexico, South Africa, Poland) relies on shorter MSCI and local-bond series that carry wider bands, with DFBETA influence diagnostics pre-specified for the calibration regression. Sources are DMS (Dimson, Marsh and Staunton, 2002) and JST (Jordà, Knoll, Kuvshinov, Schularick and Taylor, 2019) for long-run return robustness, MSCI local-currency series for the baseline moments, Falk et al. (2018) for z_c, and OECD Pensions at a Glance and Average Annual Wages for the pension and wage inputs. The prompt stake W is the round local-currency equivalent of $50,000.

Table A13: The pre-registered hypotheses, their tests, and their outcomes

	Statement	Test and correction	Status	Result
H1	Equity recommendations differ systematically across countries (country fixed effects in Eq. (1)).	Joint wild-bootstrap Wald test of all twenty contrasts, with the per-country contrasts corrected as a single Holm family.	Primary	Supported. $F(20, 12) = 20.56$, $p = 0.006$. Seven of twenty contrasts survive the Holm correction, all negative, from Brazil (-11.0 pp) to Mexico (-2.7 pp).
H2	Advice does not move with country fundamentals (the U.S.-default anchor): $\kappa = 0$ in the calibration regression (12).	Wild bootstrap over the twenty non-U.S. countries, HC2 standard errors alongside.	Second altitude[a]	Descriptive read consistent with compression. $\hat{\kappa} = 0.13$ (HC2 se 0.11) at the baseline grid point, ranging 0.07–0.43 across the sensitivity grid and always far below one. The inference of record is the wild bootstrap over countries, which does not reject $\kappa = 0$ ($p = 0.32$) and rejects full calibration $\kappa = 1$. The two-stage bootstrap that propagates only the first-stage model-sampling uncertainty, holding the country panel fixed, bounds the slope in $[0.07, 0.19]$, but this reflects a narrower source of variation than the primary procedure and is reported as secondary. The registered takeaway is no movement with fundamentals detectable under the primary test, and at most a small positive slope far below the unit slope of full calibration (Section 5.6).
H3	The U.K. contrast is equivalent to zero within a ± 5 pp margin, consistent with shared Anglo-American norms.	Two one-sided tests: supported when the 90 percent wild-bootstrap CI for γ_{UK} lies inside the margin.	Primary	Supported. $\hat{\gamma}_{UK} = +0.7$ pp with 95 percent CI $[-0.8, +2.2]$, so the stricter 90 percent interval is inside the margin as well.
H4	Asking in the local language rather than in English shifts advice, holding the country name fixed ($\lambda \neq 0$ in Eq. (3)).	Single pre-specified coefficient, reported uncorrected.	Primary	Supported. $\hat{\lambda} = -1.3$ pp, $p = 0.007$, CI $[-2.3, -0.4]$.
H5	Naming local instruments and pension vehicles shifts advice beyond the country name alone ($\phi \neq 0$ in Eq. (2)).	Single pre-specified coefficient, reported uncorrected.	Primary	Not supported at the 5 percent level. $\hat{\phi} = -0.7$ pp, $p = 0.069$, CI $[-1.5, +0.1]$.
H6	The age-55 tilt toward bonds is smaller where pension replacement rates are higher (age-55 $\times$ standardized replacement rate in Eq. (4)).	Single pre-specified interaction, reported uncorrected.	Secondary	Not supported. -0.02 pp, $p = 0.964$. The coarser above-median split gives $+0.4$ pp, $p = 0.489$.
H7	Country effects covary positively with the country's GPS risk-taking score.	Bivariate slope over the twenty non-U.S. countries, wild bootstrap over countries.	Second altitude[a]	Descriptive read unsupported. Slope $+0.028$ (HC2 se 0.021), wild $p = 0.272$.
H8	Country effects covary positively with the country's financial literacy score.	Bivariate slope over the twenty non-U.S. countries, wild bootstrap over countries.	Second altitude[a]	Descriptive read unsupported. Slope $+0.055$ (HC2 se 0.040), wild $p = 0.178$.
H9	The rate at which the explanation text invokes a given rationale differs across countries (Section 7).	Country fixed effects per admitted category, one Holm family per category.	Second altitude[b]	Supported where the design gives it variation. Local-context references are elevated in all twenty non-U.S. countries ($+30$ to $+78$ pp, nineteen of twenty Holm-surviving) and inflation reasoning is most elevated in the five high-inflation countries (Turkey $+56$ pp, South Africa $+33$ pp, both Holm-surviving). Pension-as-exposure rates are flat everywhere (all $\lvert$estimates$\rvert \leq 8.5$ pp, none survive). Horizon

Figure A1: Age glide path by country (Question 1, Version B)



Notes: Mean recommended equity share by stated investor age (30, 40, or 55) within each country, Version B of Question 1, with 95 percent confidence intervals on the cell means. The downward glide path has nearly the same level and slope in every country, the raw pattern behind the null interaction of age with pension generosity in equation (4).

B Technical Appendix: Translation, Collection, and Parsing

This appendix documents the execution machinery behind Sections 3 and 3.7: the translation pipeline and its quality controls, the collection infrastructure, and the parsing, refusal-detection, and validation steps. The replication package contains the pipeline code, the full prompt texts for every country, version, and question, and the translation review and provenance records.

B.1 Prompt Structure and Version Design

Each prompt consists of two blocks. The *Instructions* block specifies the JSON schema and the word limit on the explanation and is always written in English, regardless of the country/language of the rest of the prompt. The *Question* block contains the investor's stated demographics and the financial question itself and is the only part of the prompt that varies across versions. Keeping the Instructions block in English is a deliberate design choice validated in pilot: it ensures that JSON parsing succeeds at near-identical rates across all fourteen languages (thirteen non-English languages plus English, with Swiss German and Mexican Spanish as within-language variants) and removes a potential confound between language treatment and response-format compliance.

B.2 Translation pipeline and quality control

Version B and Version C question blocks were produced with the DeepL API through an automated pipeline, while Version B-cross is the original English prompt and required no translation. The pipeline translates only the Question block, never the Instructions block, protects the pinned financial terms and the locked currency amounts through the API's term-protection markup so that they survive translation verbatim, requests the formal register where the target language supports it, and appends every translation call with its raw response to a provenance log. Ages were translated as concrete values rather than placeholders after live testing showed that a protected {`age`} placeholder derailed the translation (a Swedish block came back reading "my name is {age} years"). The per-age Question 1 blocks were then collapsed, after all corrections had been applied, to a single canonical block per country and version with the age digit substituted mechanically at query time, so that the age-30, age-40, and age-55 prompts within a cell differ only in the stated age.

Machine output passed through two automated checks with documented failure modes. A pin-survival check verifies that every protected term and amount appears verbatim in the translated block, and it caught three systematic failures in which the translator dropped a locked currency amount, rendered an amount as the name of a financial product, or dropped a pension term. A complete back-translation of every unique question block to English was scored with smoothed sentence BLEU against the source, with 0.6 as the flagging threshold. The round-trip score proved too weak to serve as an acceptance gate.

It penalizes faithful translations into linguistically distant languages, flagging every Japanese, Korean, and Chinese block, and it passed at least one genuinely wrong translation, so it was used to route blocks to human review rather than to clear them.

The operative gate was native-speaker review of the Version B and Version C blocks in all fifteen translated countries, conducted in waves and prioritized where the financial vocabulary diverges most from the generic European templates (Chinese, Korean, Japanese, Turkish, and Brazilian Portuguese), where machine translation is weakest (Finnish, the panel's only Finno-Ugric language), and for the Swiss German and Mexican Spanish variants, with later waves covering the remaining European languages, Swedish (checked by the author, a native speaker), and the local-instrument wording of the English-prompting India cell. South Africa, the one country with no available native reviewer, was instead verified through a documented two-vendor LLM cross-check at temperature zero with a pre-stated escalation rule, which surfaced one instrument-naming error that was corrected through a dated design-lock amendment and re-verified on both vendors. The same cross-check was run on the UK, Australia, and Canada blocks and re-run on Mexico and Spain after their native reviews, with vendor suggestions adopted only where they did not contradict a documented native-reviewer decision. Reviewer corrections were not applied by hand to the translated files. They were recorded in a persistent override layer that the pipeline re-applies after any re-run, with the term-protection check re-executed on every corrected block, so that a later re-translation cannot silently overwrite a reviewer's fix. Before the pre-registration was filed, the complete grid passed a deterministic consistency audit of every assembled prompt with zero failures, an adversarial model-based review of all ninety unique translated blocks, and a full human read of every block, and the translations were then frozen for collection.

B.3 Collection infrastructure

The collection runner submits queries through the OpenRouter API with the design parameters of Section 3 and is built so that an interruption cannot produce a duplicate analysis record. Two append-only logs record every query. The parsed results file is the leading log: for each response the runner writes and flushes the parsed row first and the raw API response afterwards. On restart the runner reads the identifiers of completed queries from the parsed results file and submits only the remainder, so no already-recorded observation is ever queried twice. Two edge cases remain and are benign. A crash after a response returns but before its parsed row is flushed leaves that single in-flight query unrecorded, so it is re-queried on restart, at most one repeated API call per interruption. A crash between the parsed-row write and the raw-response write leaves that one recorded observation without its raw audit copy, without affecting the parsed data used in the analysis. Requests were paced below provider rate limits, retried automatically in a way that honors the provider's retry-after signal, and distributed across worker threads at a concurrency validated in throughput probes that produced no rate-limit rejections.

The main run finished with exact coverage of the design grid, every cell reaching its fifty repetitions with no duplicates and no unresolved API errors.

B.4 Parsing, refusal detection, and validation

Parsing proceeds in a fixed order of decreasing strictness. The parser first extracts the first balanced brace block in the response and attempts to read it as JSON against the schema stated in the Instructions block. If that fails, it searches for a decimal fraction, and then for a percentage, in each case only where the number appears in explicit allocation language, so that a stray number elsewhere in the response cannot be mistaken for a recommendation. The strategy that produced each observation is recorded in a parse-method field, which allows any analysis to be re-run on the strictly JSON-parsed subsample. Responses whose generation stopped at the 4,000-token ceiling are marked truncated and excluded under the pre-registered rule.

Refusal detection runs after parsing, not before. A response is classified as a refusal only when a refusal phrase matches and no numeric recommendation was parsed, which prevents advisory boilerplate such as "consult a financial advisor" on an otherwise valid answer from being misclassified. The phrase list covers all fourteen prompt languages (thirteen non-English languages plus English, with Swiss German and Mexican Spanish handled as within-language variants) and was extended at translation time as each language entered the design. The main collection produced no refusals.

Model identity is monitored at the level of the individual response. The dated model versions behind each requested identifier were pinned before launch, and every row records both the requested and the returned identifier. In the main collection, every mismatch between the two was the provider's resolution of a version alias to its dated snapshot, a one-to-one mapping that was stable across the collection window, so no model rotated versions mid-run. Staging and validation precede any estimation. A staging script reduces the raw collection to the analysis columns and verifies the absence of duplicate query identifiers, and a validation stage recomputes cell counts against the design grid, parse and refusal rates by cell, and the requested-to-returned identifier map before the first regression runs.

B.5 Benchmark I: pension-wealth and human-capital inputs

The two non-tradable balance-sheet inputs of Benchmark I (Section 5.3) have simple annuity expressions. Pension wealth for an average earner is the replacement-rate-scaled wage stream from retirement age R_c for the remaining life span L_c, discounted to age 40,

$$P_c \;=\; \rho_c\, \bar{w}_c\, (1+g_c)^{R_c-40}\, \frac{1-(1+r_c)^{-L_c}}{r_c}\, (1+r_c)^{-(R_c-40)}, \tag{15}$$

where ρ_c is the net pension replacement rate and $\bar{w}_c$ the average wage. The wage-growth factor keeps P_c consistent with equation (16), which assumes that replacement rates are defined relative to pre-retirement earnings rather than today's wage. Three further assumptions are that the annuity is an ordinary level annuity over L_c years whose factor tends to L_c as r_c approaches zero (relevant for Japan), that survival weighting is absorbed into L_c, and that the net replacement rate must be paired with a net wage measure when the calibration table is filled in.

Human capital is the discounted wage stream from now to retirement,

$$H_c \;=\; \sum_{t=1}^{R_c-40} \bar{w}_c \,(1+g_c)^t \,(1+r_c)^{-t}, \tag{16}$$

with g_c real wage growth.